Multiple dimensions of semantic and perceptual similarity contribute to mnemonic discrimination for pictures

Article (Accepted Version)

Multiple dimensions of semantic and perceptual similarity contribute to mnemonic discrimination for pictures

Loris Naspi[1], Paul Hoffman[1], Barry Devereux[2], Tobias Thejll-Madsen[1], Alex Doumas[1], and Alexa Morcom[3]

[1] School of Philosophy, Psychology and Language Sciences, University of Edinburgh

[2] School of Electronics, Electrical Engineering and Computer Science, Queen's University Belfast

[3] School of Psychology, University of Sussex

Abstract

People often misrecognize objects that are similar to those they have previously encountered. These mnemonic discrimination errors are attributed to shared memory representations (gist) typically characterized in terms of meaning. In two experiments, we investigated multiple semantic and perceptual relations that may contribute: at the concept level, a feature-based measure of concept confusability quantified each concept's tendency to activate other similar concepts via shared features; at the item level, rated item exemplarity indexed the degree to which the specific depicted objects activated their particular concepts. We also measured visual confusability over items using a computational model of vision, HMax, and an index of color confusability. Participants studied single (Experiment 1, N = 60) or multiple (Experiment 2, N = 60) objects for each basic-level concept, followed by a recognition memory test including studied items, similar lures, and novel items. People were less likely to recognize studied items with high concept confusability, and less likely to falsely recognize their lures. This points to weaker basic-level semantic gist representations for objects with more confusable concepts because of greater emphasis on coarse processing of shared features relative to fine-grained processing of individual concepts. In contrast, people were more likely to misrecognize lures that were better exemplars of their concept, suggesting that enhanced basic-level semantic gist processing increased errors due to gist across items. False recognition was also more frequent for more visually confusable lures. The results implicate semantic similarity at multiple levels and highlight the importance of perceptual as well as semantic relations.

Keywords: mnemonic discrimination, semantic gist, perceptual gist, episodic memory, recognition memory

Introduction

Memory for unique experiences depends on the ability to discriminate between events that share multiple overlapping features. For example, we may misrecognize an unknown car in a car park as one we have just hired because it is similar in type and color. Memory theory explains these mnemonic discrimination errors in terms of a gist shared between incoming information and previously encoded representations (Reyna & Brainerd, 1995). Gist is assumed to embody essential meaning shared by multiple items, but the representations underpinning gist are poorly understood. Mnemonic discrimination errors are observed for a wide range of materials including pictures, words, and narratives (Brainerd & Reyna, 2002). More direct evidence also suggests that multiple semantic relations may contribute to the tendency to falsely recognize similar items as having been studied (Brainerd et al., 2008; Cann et al., 2011; Coane et al., 2016; Montefinese et al., 2015). Moreover, although gist is typically conceptualized as semantic, shared perceptual information may also be important (Koutstaal et al., 2003; Pidgeon & Morcom, 2014). Here, we combined objective measures of semantic and perceptual similarity with a linear mixed modelling approach to tease apart multiple influences on mnemonic discrimination in one exploratory study and one confirmatory, preregistered study.

Studies using categorized pictures suggest that semantic relations at different organizational levels may impact mnemonic discrimination. In the typical task, individuals study pictures of multiple exemplars of basic-level concepts (e.g., several cats). At test, memory probes include studied items (e.g., the same cat) together with lure items belonging to the same studied basic-level concepts (e.g., a different cat), as well as novel items that do not belong to any of the studied concepts (e.g., a snake) (Koustaal et al., 1999; Koutstaal & Schacter, 1997). Participants must decide whether or not they have previously been shown each picture. Koutstaal and Schacter (1997) found that people were about 24% more likely to falsely endorse the related lures as "old" than they were to endorse unrelated novel items, and subsequent studies have found a similar pattern. Intuitively, studied items and lures (e.g., a different cat) tend to be semantically as well as perceptually similar. Koutstaal and Schacter (1997) attributed these memory errors to processing of semantic and/or perceptual gist. Without specifying the relative roles of these kinds of relations, gist has been proposed to reflect properties like similarity that are shared between studied exemplars and lures (e.g., Gutchess & Schacter, 2012; Koutstaal et al., 2003; Pidgeon & Morcom, 2014; Slotnick & Schacter, 2004).

Variants of the categorized picture task have also been used in which studied and lure objects were related at the superordinate-category level (e.g., land animals: a cat, a horse, a cow), rather than at the basic level. In these tasks, lures in the recognition memory test (e.g., a different land animal, such as a lion) are related to studied items by membership of the same superordinate category (Bowman et al., 2019; Seamon et al., 2000). As for the basic-level version of the task, false recognition is more frequent to lures than to unrelated novel items. The findings from both versions of the categorized pictures task suggest that mnemonic discrimination of pictures

depends on multiple semantic as well as perceptual relations, although these studies could not distinguish the different influences. Such results are also consistent with studies using verbal materials in which studied items and lures are semantically related either by superordinate category (Brainerd et al., 1995) or at a higher level such as a narrative (Reyna et al., 2016).

According to fuzzy trace theory, lures elicit false recognition errors when a gist memory representation is not opposed by a detailed memory representation of the related studied items (Reyna & Brainerd, 1995). The gist memory is an episodic trace that represents meanings and relations shared by sets of events, but the informational content of these relations is not specified (Brainerd et al., 2008; Roediger et al., 2001). In the Deese-Roediger-McDermott paradigm (DRM) the studied word lists (e.g., bed, rest, awake) are all associated to a non-presented critical lure (e.g., sleep) (Deese, 1959; Roediger & McDermott, 1995). Frequent false recognition of the critical lures is attributed to these backward associations rather than to shared meaning. However, although critical lures are not typically semantically similar to most studied list items (e.g., sleep is not very similar to bed), they do share other semantic relations (e.g., familiarity and meaningfulness; Brainerd et al., 2008). Preliminary data suggest that lure words that are highly similar to studied items or share thematic information may elicit additional errors that cannot be explained by associative strength alone (Cann et al., 2011; Coane et al., 2016; Montefinese et al., 2015). Coane et al. (2016) showed that people were more likely to misrecognize lure words when studied lists shared semantic features and category membership with the lures as well as being associatively related. However, the between-list comparison could not identify specific effects of feature similarity or category membership. Montefinese et al. (2015) more directly investigated the effects of feature similarity using an index of the number of shared semantic features between pairs of concepts derived from norms for production frequency (McRae et al., 2005; Montefinese et al., 2013). After studying sets of categorically related words (e.g., car, bicycle, truck), participants were more likely to falsely endorse as "old" unstudied items that shared more semantic features to their studied items (e.g., bus compared to plane) (see also Montefinese et al., 2018). However, as there was no baseline novel item condition, it is unknown whether the effects of feature similarity on false alarms reflected a real effect on memory or a modulation of response bias.

There is less agreement about the role of perceptual similarity in mnemonic discrimination. While fuzzy trace theory characterizes false recognition of similar lures in terms of semantic gist, other theories explain it in terms of generic similarity and therefore also predict errors due to perceptual relations between studied items and lures. In global matching models, false recognition (like true recognition) reflects feature overlap with stored memory traces, including visual context as well as semantic relations (Arndt & Hirshman, 1998; Arndt, 2010). Likewise, pattern separation/completion accounts describe mnemonic discrimination in terms of complementary computational processes that act to minimize overlap between new and existing memory traces along multiple dimensions of similarity, and to reinstate stored traces in response to partial cues at test (Wilson et al., 2006; Yassa & Stark, 2011). Several studies using verbal material have suggested that perceptually-driven errors can occur when lures rhyme with lists of studied rhyming words (Budson et al., 2003; Watson et al., 2003; Watson et al., 2001). However, in the majority of studies, which have presented words visually (for an

exception see Reyna & Kiernan, 1994), these errors cannot be attributed simply to sensory properties of the stimuli but may also reflect similarity of phonological and/or orthographic representations. Further evidence for perceptual influences on mnemonic discrimination comes from studies showing effects of shared perceptual context in the form of distinctive fonts (Arndt & Reder, 2003). Such effects have also been demonstrated when lure words are not also semantically associated with studied items (Burnside et al., 2017).

The effects of semantic and perceptual relations are more difficult to separate for pictorial material, because some semantic features of objects can be directly perceived (e.g., it can be seen from an image whether an object < has legs >, < is red >). To isolate the effects of semantic similarity on mnemonic discrimination of pictures, Koutstaal et al. (2003) compared memory for sets of colored drawings of abstract and concrete objects. While concrete images depicted meaningful objects that were exemplars of the same basic-level concept, as in the standard categorized pictures task, the sets of abstract images were created to be pre-experimentally meaningless but perceptually similar. False recognition was about 10% higher for lures related to studied concrete than abstract categories, supporting a specific role for semantic as opposed to perceptual gist (for a replication, see Pidgeon & Morcom, 2014). Perceptual effects were suggested by above-zero false recognition of abstract objects, although this conclusion assumed that the perceptual relatedness of the concrete and abstract objects was matched and that the abstract objects were not processed semantically. This second assumption was questioned by subjective reports from Pidgeon and Morcom (2014)'s participants of spontaneous verbal labelling of the abstract images. Therefore, this paradigm could not unambiguously identify distinct semantic and perceptual contributions to mnemonic discrimination errors. Perceptual effects are suggested by findings that people make errors to lures that are rotated photographs of studied stimuli, in old/new (Motley & Kirwan, 2012) as well as forced-choice (Brady et al., 2008) recognition tasks. In such cases the task is to identify the same image, so in that sense the lures are related semantically to the studied object and the two elements difficult to separate. However, increasing the perceptual (rotational) difference does reduce lure errors (Motley & Kirwan, 2012). There is also recent evidence of perceptual-level interference in true recognition in a retrieval-induced forgetting task as a result of similarity of object shape and color (Reppa et al., 2020). In these studies only one dimension is varied at a time, so perceptual and semantic variables have not been shown to influence memory for the same items. Konkle et al. (2010) separated these elements more objectively, with semantic and perceptual distinctiveness ratings in a task using large sets of pictures organized at the basic level. They found that mnemonic discrimination errors in a forced choice recognition task were slightly more frequent for larger sets of objects that had been rated as less semantically distinctive, but perceptual effects were not significant. Their two ratings were relatively uncorrelated, although it cannot be assumed that perceptual properties did not influence semantic ratings or vice versa (see also Pidgeon & Morcom, 2014).

The Current Research

The studies reviewed above have established that semantic relations between studied and unstudied items are a key driver of mnemonic discrimination errors, but it remains unclear what kind of semantic information is critical. Moreover, although gist is usually conceptualized as semantic, perceptual similarity may also be important, particularly for rich pictorial material. The two experiments reported here used objectively quantified dimensions to investigate the effects of semantic and perceptual similarity on mnemonic discrimination in a typical categorized pictures task. In this task, items are pictures of individual objects, so lures are different exemplars of the same basic-level concept. We operationalized semantic similarity at two distinct organizational levels: at the *concept level*, indexing the relations between the basic-level concepts, and at the *item level*, indexing the relations between individual exemplars and their basic-level concepts. Perceptual similarity was also quantified using properties shared between items.

We used feature overlap to measure semantic similarity at the concept level. A large body of experimental evidence supports the view that semantic memory is structured in terms of features. According to distributed feature models, like the *Conceptual Structure Account* (CSA; Tyler & Moss, 2001), concepts are represented in the brain by their features (e.g., < has legs >, < has eyes >, < has a tail >) in a connectionist system in which the mutual co-activation of the feature nodes determines the semantic processing (McRae et al., 1997; Tyler & Moss, 2001; Vigliocco et al., 2004). The statistical regularities of semantic features, derived from property norms, have proven to be a useful way of characterizing the structure and content of semantic representations (Devereux et al., 2014; Garrard et al., 2001). The most prominent statistical characteristic assumed to structure the semantic space and determine how concepts are processed is concept confusability (Clarke & Tyler, 2014). Concept confusability measures the degree to which a concept's semantic features are shared with other concepts (e.g., many animals < have ears >). Highly shared features (e.g., < has legs >, < has eyes >, < has a tail >) provide coarse information about the superordinate categories to which a concept belongs (e.g., land animals), and support decisions that depend only on this coarse-grained information. In contrast, accessing specific basic-level concepts (e.g., tiger) requires finer-grained semantic processing that includes features more distinctive to the particular concept (e.g. < has stripes >), and support more specific tasks like naming. In the current work, we used concept confusability as an index of information shared across the basic-level concepts.

To complement the examination of concept confusability, we assessed effects of semantic and perceptual relations at the item level, via properties of the individual depicted exemplars. Such properties are likely to be important determinants of people's ability to correctly reject lures which are unstudied exemplars of studied basic-level concepts. Previous studies have shown that basic-level processing is enhanced for objects that are more representative of a stored basic-level conceptual representation. Pictures that are better exemplars are categorized faster at the basic level than pictures that are poorer exemplars (Barry et al., 1997; Snodgrass & Vanderwart, 1980). Therefore, we used ratings of item exemplarity to index how well a picture corresponded to its basic-level representation. We assumed that similarity between a picture

and its conceptual representation would facilitate gist processing at the item level, i.e., information shared across exemplars of each concept.

To operationalize perceptual similarity, we drew on an established computational model of perceptual processing. The *Hierarchical Model and X* (HMax; Riesenhuber & Poggio, 1999; Serre et al., 2007) models different hierarchical stages of the ventral processing stream in different layers, progressing from early visual cortex (V1) to posterior inferior temporal cortex (IT). The C1 layers correspond to increasingly position- and scale-invariant early visual cortex (V1/V2) which maintain feature specificity, while C2 layers simulate the extrastriate visual area cells (V4/IT) that integrate visual features from previous layers to represent object shape. Measures based on these two layers have been validated in studies of visual object recognition that have distinguished the time courses and neural correlates of semantic versus visual processing (Clarke & Tyler, 2014; Clarke et al., 2015). Here, we generated measures of the visual confusability of each image with others in the set, defined in terms of the properties indexed by C1 and C2 (e.g., orientation and shape). Lastly, since the HMax model does not represent color, we computed a novel index of color confusability using the CIELab color space, known to be an approximation of human color perception (Rubner et al., 2000).

In a first, exploratory, experiment, we examined mnemonic discrimination in a simple task in which participants studied one exemplar for each basic-level concept and were later tested on a single lure (Bakker et al., 2008; Lacy et al., 2011; Reagh et al., 2016; Stark et al., 2013; Yassa et al., 2011). Then in a second, preregistered, confirmatory experiment, we sought to replicate and explain the findings of the first by increasing the number of studied exemplars of each basic-level concept. Successful mnemonic discrimination requires accurately identifying studied items and rejecting similar lures. Since distinct processes may contribute to memory for these two trial types, performance is typically assessed using separate measures of true and false recognition, and many studies focus mainly on false recognition. To adjust for response criterion, sensitivity measures derived from signal detection theory can be computed from rates of endorsement of studied and lure items as "old", relative to unrelated unstudied items. Equivalent measures of studied relative to lure item endorsement can also assess overall ability to discriminate studied items from lures in memory. In both experiments, we used generalized linear mixed effects models, which provide parameter estimates to index participants' memory sensitivity and response bias that are equivalent to those provided by the equal variance signal detection theory framework (DeCarlo, 1998). This allowed us to preserve and take into account variability across items as well as participants, which results in an increased accuracy and generalizability of the parameter estimates (Baayen et al., 2008; Quené & van den Bergh, 2008).

In this task, the studied and lure exemplars of the same basic-level concept can be regarded as sharing a semantic gist and possibly also a perceptual gist across items. We reasoned that concept confusability would impact mnemonic discrimination by weakening the basic-level conceptual representations shared by studied items and their lures. If a concept is highly confusable with other concepts because they share semantic features, the concept's distinctive, basic-level representation will be less likely to be encoded. Thus, we expected that studied

objects whose concepts are more confusable would be less likely to be recognized and their corresponding lures more likely to be successfully rejected. For example, if a cherry is studied – a concept with high confusability due to multiple highly shared features like < does grow on trees >, < is sweet >, < is edible > – a weak representation of "cherry" will be encoded alongside the details of the particular exemplar. In contrast, if a foot is studied – a concept with low confusability due to its few shared features and its distinctive features < has toes >, < is found at the end of a leg > – a strong representation of "foot" will be more likely to be encoded. In a memory test, the same exemplar of "cherry" will thus be less likely to be recognized than the same exemplar of "foot", since memory for the former is weaker. For the same reason, for lures, a different exemplar of "cherry" will be more likely to be successfully rejected than a different exemplar of "foot", given weaker encoding of the studied concept. In terms of the item-level metrics, we assumed that objects with higher rated item exemplarity would more strongly engage the basic-level conceptual processing shared by studied items and lures. We therefore predicted that higher exemplarity lures would be more likely to be misrecognized. Lastly, based on similar logic we predicted more frequent false recognition of more visually confusable lures.

Materials and Methods

Participants

The study included sixty participants aged 18-33 years (M = 21; SD = 2.2; 15 male, 45 female). Eleven further participants were excluded from data analysis: 7 due to errors in stimulus lists or data acquisition issues, 1 due to misunderstanding of instructions, and 3 who did not meet the inclusion criterion for English fluency. The sample size was determined *a priori* using the simR package in R (Version 1.0.4; Green & Macleod, 2016). We powered the study for an interaction of condition (lure vs. new) × concept confusability, based on a pilot study suggesting an effect size equivalent to Cohen's d = 0.17 (OR = 1.35). With N = 60 we had 87.30 % power to detect such an effect at alpha = .05. Inclusion criteria were fluency in English (spoken since at least the age of 5 years), and normal or corrected-to-normal vision. Data from an additional group of older participants collected at the same time will be included in a separate report. Participants were compensated financially or with course credits. They were contacted by local advertisement and provided informed consent. The study was approved by the University of Edinburgh Psychology Research Ethics Committee (Ref. 278-1617/1).

Stimuli

Stimuli were pictures of objects corresponding to 180 of the 638 basic-level concepts in the Centre for Speech, Language and the Brain property norms (the CSLB norms; Devereux et al., 2014). These 180 basic-level concepts comprised 9 members of each of 20 different superordinate categories (*Appliance*, *Bird*, *Body Part*, *Clothing*, *Container*, *Drink*, *Flower*, *Food*, *Fruit*, *Furniture*, *Invertebrate*, *Kitchenware*, *Land Animal*, *Music*, *Sea Creature*, *Tool*, *Toy*, *Vegetable*, *Vehicle*, *Weapon*) and half were living and half non-living. We sourced two images for each basic-level concept. One set of 180 was a subset of images used by Clarke and Tyler (2014), and the other was compiled from the Bank of Standardized Stimuli (BOSS; Brodeur et al., 2014) and from the Internet. Each study list comprised 120 images of exemplars

of different basic-level concepts, selected evenly from the superordinate categories (i.e., 6 different basic-level concepts per superordinate category). Each test list consisted of 180 items: 60 studied images, 60 similar lures (i.e., novel images corresponding to the other 60 studied basic-level concepts), and 60 novel items (i.e., novel images of basic-level concepts that had not been studied). Three filler trials prefaced both study and test phases. We generated 6 different study and test lists which fully counterbalanced the allocation of the basic-level concepts and the two sets of images to conditions (studied, lure, and novel).

Procedure

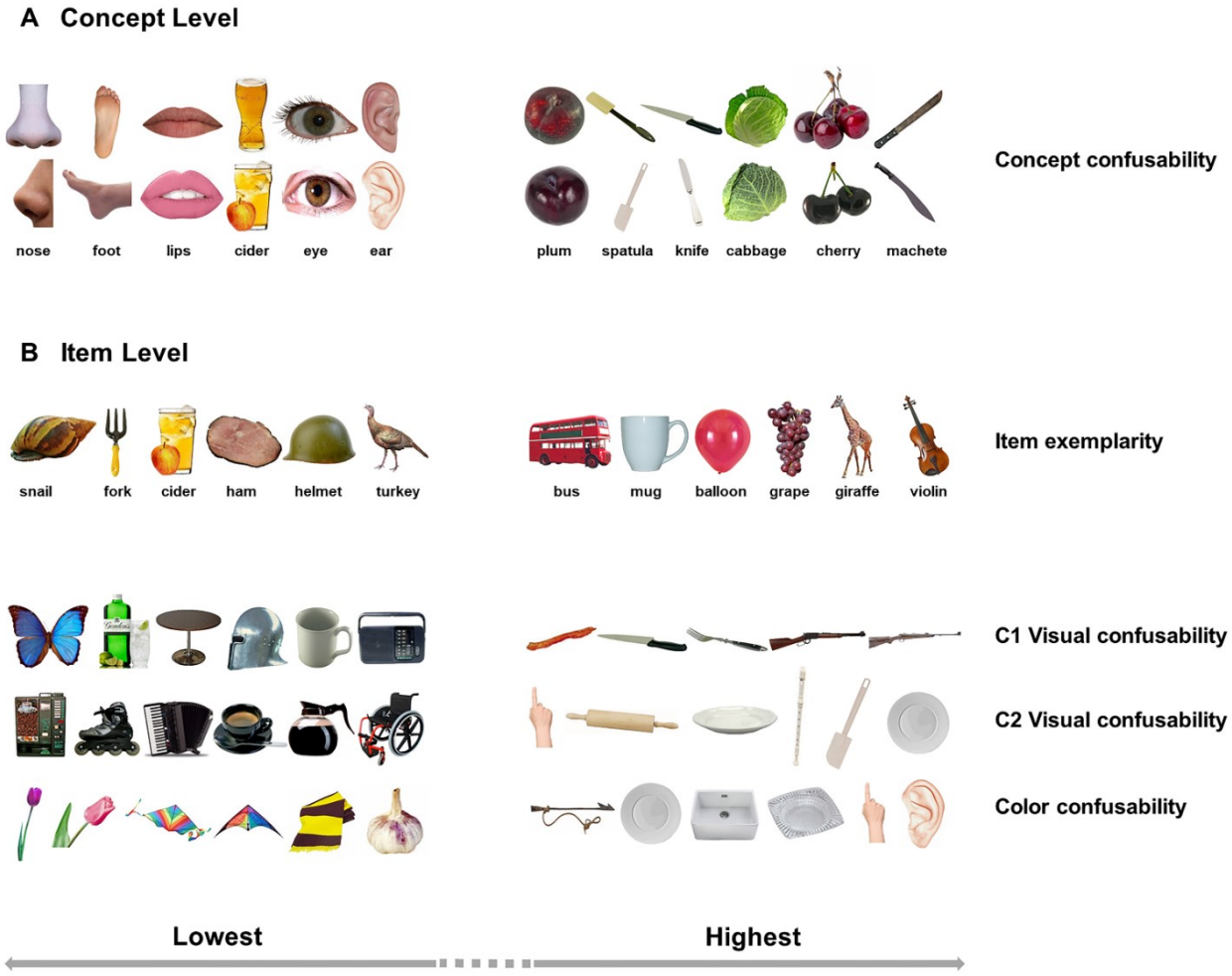
The experiment consisted of a single study phase followed by a recognition test phase. Between study and test phases, participants completed standardized cognitive tests (not reported here) for 15 minutes. Stimuli were presented using E-Prime 2.0 software (Psychology Software Tools, Pittsburgh, PA). Participants were tested individually in the laboratory. At study, they viewed one image at a time, and rated how pleasant it was from 1 (very unpleasant) to 5 (very pleasant). Participants were not informed of a later memory test. Images were presented against a white background within a 15 × 13 cm area in the center of a computer screen, and viewed at a distance of approximately 50 cm. Trials were self-paced. At test, participants viewed one image at a time every 3 s and judged each as “old” or “new” using the keyboard, indicating at the same time whether they were confident in this judgement. Responses were made using the “Z” and “A” keys when judging an item as “old” with high and low confidence, respectively; the “M” and “K” keys when judging an item as “new” with high and low confidence, respectively, and mappings of responses to hands counterbalanced over conditions. After completing the test phase, each participant also completed the concept familiarity or the item exemplarity ratings. The entire procedure took approximately 50 minutes and participants were debriefed at the end of the experiment.

Variables of Interest

The concept- and-item level variables of interest are illustrated in Figure 1.

Figure 1

Schematic Depiction of Stimuli Used in Both Experiments



Note. Rows show individual exemplars with the highest (right side) and lowest (left side) scores on each concept- and-item level experimental variable. Panel A illustrates concept confusability: the basic-level concept name is given alongside images of both exemplars representing it in Experiment 1. High confusability concepts share more semantic features with other concepts. Panel B illustrates item-level measures for individual images of objects. Item exemplarity is an index of similarity between the depicted exemplar and the concept representation. The perceptual measures define confusability of an item as the similarity with its most similar neighbor in the set. C1 and C2 were obtained from gray-scaled version of the images depicted in Figure 1. For definitions see Variables of Interest section.

Concept Level

Concept confusability. Our measure of concept confusability was based on that of Clarke and Tyler (2014), but we used a gentler weighting system and an updated set of property norms (Devereux et al., 2014). The property norms provide a matrix of features associated with 638 objects (e.g., has 4 legs, has stripes and lives in Africa are features of a zebra). These were collected by presenting participants with a written concept name and asking them to produce properties of the concept. Taxonomic features (e.g., < is a bird >) were excluded as they refer to a superordinate category and are not normally regarded as true semantic features (Taylor et al., 2012). Based on the feature norms, each object can be represented by a binary vector. Semantic similarity between concepts was computed as the cosine angle between feature vectors in a semantic feature matrix in which each concept was represented by a binary vector indicating whether each feature was associated with the concept (1) or not (0). Concept confusability with all the other concepts in the set was then calculated by a weighted sum of the similarities in which each weight was the between-concept similarity itself, i.e., the sum of squared similarities (see Figure 1). This measure emphasized feature sharedness of each concept with those concepts with which it shared many semantic features. We also calculated the number of features for each concept from the same norms.

Concept familiarity. Thirty of the 60 participants judged the 360 pictures representing the 180 concepts. Following Snodgrass and Vanderwart (1980), we asked subjects to judge the familiarity of each picture *"according to how usual or unusual the object is in your realm of experience"*. Concept familiarity was defined as *"the degree to which you come in contact with or think about the concept"*. Participants were required to give their response on a 7-point scale (1 = completely unfamiliar, 7 = completely familiar) using the keyboard. For each picture, familiarity scores were averaged over all participants (see Supplemental Materials for its theoretical relevance).

Item Level

Item exemplarity. To obtain a measure of item exemplarity we used a rating task in which pictures were presented together with their verbal label. Following Taylor et al. (2012), we asked participants to judge "*how closely each picture resembles your mental image of the object*", giving their response on a 7-point scale (1 = poor picture of concept word, 7 = excellent picture of concept word). The 30 participants who did not provide familiarity ratings were asked to provide these item exemplarity scores, which were averaged to give a single score per picture (see Figure 1).

Perceptual confusability. Measures capturing the low- and high-level visual attributes of each picture were derived for each of the 360 images (2 per concept) used to represent our basic-level concepts. We calculated two indices for each dimension which embodied two alternative hypotheses about how perceptual similarity between images might affect mnemonic discrimination. Although the second measure was preregistered for Experiment 2, for clarity we only report the results from the first in the main paper, since it gave a coherent picture across the two experiments (see also Discussion). We first extracted HMax estimates of low- and high-level visual object information: a C1 response related to early visual cortex (V1/V2), and a C2 response related to V4/posterior IT. Response vectors from the C1 and C2 layers were computed for grey-scaled versions of each image. Similarity between pictures was then

calculated using the Pearson correlation coefficients between vectors. For the main visual confusability measure, for each image we defined confusability as the similarity value with its most similar picture (i.e., the nearest neighbor; see Figure 1). The second metric, graded visual confusability, was analogous to concept confusability, indexing an image's similarity to the full set of images (see Supplemental Figure 4). We calculated a weighted score for each image by summing the squared ranks of Pearson correlations between each image i and all other images j, so pictures with high graded visual confusability scores were those that were similar to many other pictures. To obtain the ranks of all the similarity values, we first extracted the matrix of Pearson correlations between images as a single vector. Then, we computed the corresponding rank for each correlation, and transformed the vector of ranks back into a matrix. This allowed us to deal with negative Pearson correlations, which were assigned the lowest ranks.

For each image, we also generated a nearest neighbor index of color confusability using the color distance package (Weller & Westneat, 2019) in R (version 3.4.3; R Core Team, 2017). This measure represents the degree to which the color of each item resembled that of the most similarly-colored item in set. After converting the RGB channels into CIELab space, we calculated the earth mover's distance between each pair of images (Rubner et al., 2000). We then normalized the distance and transformed the distance matrix in a similarity matrix using the equation S = D - 1 so that similarity values ranged from 0 (lowest similarity) to 1 (highest similarity). Then, for each item, we retained the similarity with its most similar item in the set (see Figure 1). As described above for the HMax measures, we also calculated a measure of graded color confusability. From the similarity matrix, to obtain a single metric for each image, we summed the squared ranks of similarities, so higher values indicate greater color confusability with all the other pictures in the set (see Supplemental Figure 4).

Nuisance Variables

Mnemonic discrimination may be influenced by a range of other visual, phonological, lexical, and semantic factors in addition to the semantic and perceptual confusability measures of interest here. We controlled for the effect of the following nuisance variables, described in more detail in the Supplemental Materials: forward and backward associative strength estimated using a continuous association task (De Deyne & Storms, 2008; Nelson et al., 2004), word frequency (van Heuven et al., 2014), concreteness (Brysbaert et al., 2014), age of acquisition (Brysbaert & Biemiller, 2017), phonological neighborhood density (Baayen et al., 1995), the number of non-white pixels, color entropy (Chouinard & Goodale, 2012), and concept familiarity (derived from our rating task; see also Taylor et al., 2012).

Statistical Analysis

We tested our hypotheses using a series of generalized linear mixed-effect model analyses with the function glmer from the lme4 package in R (version 1.1-17, Bates et al., 2015). The linear mixed-effect model approach affords greater robustness and generalizability of inference compared to the analysis of variance typically used in studies of memory (Baayen et al., 2008). Modelling random effects of items as well as participants is helpful in studies of memory where generalization to other stimulus sets as well as other participant samples is desirable (Clark, 1973). Accuracy was modelled using a multiple linear logistic regression on participants' binary recognition judgments ("new" = 0, "old" = 1) fitted by means of a logit link function.

Thus, rather than considering our data in terms of proportions of hits, misses, false alarms, and correct rejections, as in the standard recognition test analysis, we directly predicted behavioral outcomes from item status (i.e., studied, lure, or novel). In this way we estimated how the probability of judging an item as "old" depended on its actual status, and asked which of our variables of interest moderated this effect. The log-odds-ratio coefficients generated by the generalized linear mixed effect logit models are formally proportional to *d'* in a Gaussian signal detection analysis (*d'* ≈ .6 logOR; DeCarlo, 1998; Wright et al., 2009). This relation holds over a wide range of values, so logit and probit models yield equivalent results except at the extremes. In the Results section we therefore report effect sizes in terms of *d'* equivalent for ease of comparison with other studies. We also performed an exploratory linear mixed-effect analysis of response times (RTs) at study to check whether the experimental variables would impact (non-speeded) decisions about item pleasantness. The results did not reveal any significant effects, and are not reported further (but can be found on https://osf.io/ndk83/).

To test specific predictions about memory for studied items and lures, we set the reference level for the condition factor to "novel" so that with simple contrasts we could examine modulations of a) the probability of correctly identifying studied items as "old" relative to novel items (an index of sensitivity for studied items equivalent to *d'*, reflecting true memory), b) the probability of misrecognizing related lures relative to novel items (an index of sensitivity for related lures equivalent to *d'*, reflecting false memory). This also yielded c) the probability of falsely judging novel items as "old" (an index of baseline false alarms, equivalent to the response criterion *c*). So for example, an estimated *d'* = 1.5 for an interaction between a continuous variable and the contrast of lure versus novel items, means that a one SD increase in that continuous variable is associated with a 1.5 x increase in *d'* for lures versus novel items. In a final set of contrasts, we set the reference level for the condition factor to "lure" to assess modulations of d) the probability of endorsing studied items as "old" relative to lures (an index of overall sensitivity equivalent to *d'*). This allowed us to evaluate the effect of the semantic and perceptual variables on overall mnemonic discrimination performance, a net effect of their modulations of true and false memory (Koutstaal & Schacter, 1997; Loiotile & Courtney, 2015).

For each model, we specified *a priori* random intercepts of both participants and concepts (Matuschek et al., 2017). In the fixed part, our variables of interest were condition (studied, lure, novel), two concept-level variables (i.e., concept confusability, number of features), and four item-level variables (i.e., item exemplarity, C1 and C2 visual confusability, color confusability). Within the concept-level and item-level partitions we also included the corresponding interactions with condition. To minimize model complexity, and because the nuisance variables were moderately-to-highly intercorrelated, we performed data reduction of the nuisance variables with principal components analysis (PCA) using the prcomp function in R. PCA with varimax rotation produced a 7 factor solution which accounted for 86.14% of variance (see Supplemental Table 5). We then compared the goodness-of-fit between i) the confounds model with the original nuisance variables and ii) the reduced confounds model with the principal components using the corrected Akaike information criterion (AICc). We also complemented the goodness-of-fit measure provided by AICc with the corresponding Bayesian

information criterion (BIC) and the likelihood ratio test (LRT). Model selection revealed that the simpler confounds model with principal components provided a better goodness-of-fit (AIC = 14420.23, BIC = 14492.97) relative to the model with the original nuisance variables (AIC = 14424.65, BIC = 14519.21; for LRT for a difference between models, $\chi2(3) = 1.59$, $p = 0.662$). Thus, these 7 principal components were included as nuisance variables for our models.

Model selection was carried out to determine the fixed effects structure with the best goodness-of-fit based on sets of theoretically motivated predictors, which included concept-level, item-level, and confound principal component variables. Starting with the most complex model, we used the AICc to compare progressively simpler models. At each step, we verified whether the exclusion of a particular set (in order, concept-level, item-level, and nuisance variables) was justified or not. We also supplemented this measure with the corresponding BIC and the LRT between models. Model comparison was performed using the AICcmodavg package (version 2.3-1) and the anova function in R. *Post-hoc* analyses were conducted, and results interpreted after selecting the best model. All the continuous predictor variables were standardized, and the resulting β coefficients representing log-odd-ratios were used to calculate the corresponding *d'* coefficients.

Results

The results of the mixed effects analysis for the winning model are shown in Table 1, and illustrated in Figure 2. Raw recognition responses by item type are reported in Supplemental Figure 5. The initial model comparison suggested that the full model with concept-level, item-level, and confounds principal component variables was the most parsimonious model for the data. This model received the lowest AICc score (AICc = 8821.92), indicating decisive evidence relative to simpler models (see Supplemental Table 6). It received substantial weight (AICc weight = 0.98) of the total weight of the models considered, with an evidence ratio between the top-ranked model and the second-ranked model of 58.15. That is, the evidence was 58.15 times stronger for the best model. This result was also supported by a significant LRT in favor of the full relative to the model ranked second on the basis of AICc ($\chi2(7) = 19.59$, $p = 0.007$). Unlike AICc and LRT, BIC provided evidence in favor of the model with concept-level variables only as the most parsimonious model (BIC = 8991.90). However, when we allowed the free selection of variables rather than the comparison based on our three blocks (i.e., removal of concept-level, item-level, and confounds principal component variables in this order), BIC resulted in a better goodness-of-fit when item exemplarity and C1 visual confusability were included in the model together with the concept-level variables (BIC = 8827.70). Participants' responses were collapsed across confidence judgments as results were qualitatively similar when high confidence mnemonic discrimination was analyzed. Coefficients in Table 1 represent log-odds-ratios with the corresponding *d'* effects. All the *p*-values reported below are FDR-multiple comparison corrected (Benjamini & Hochberg, 1995).

Table 1

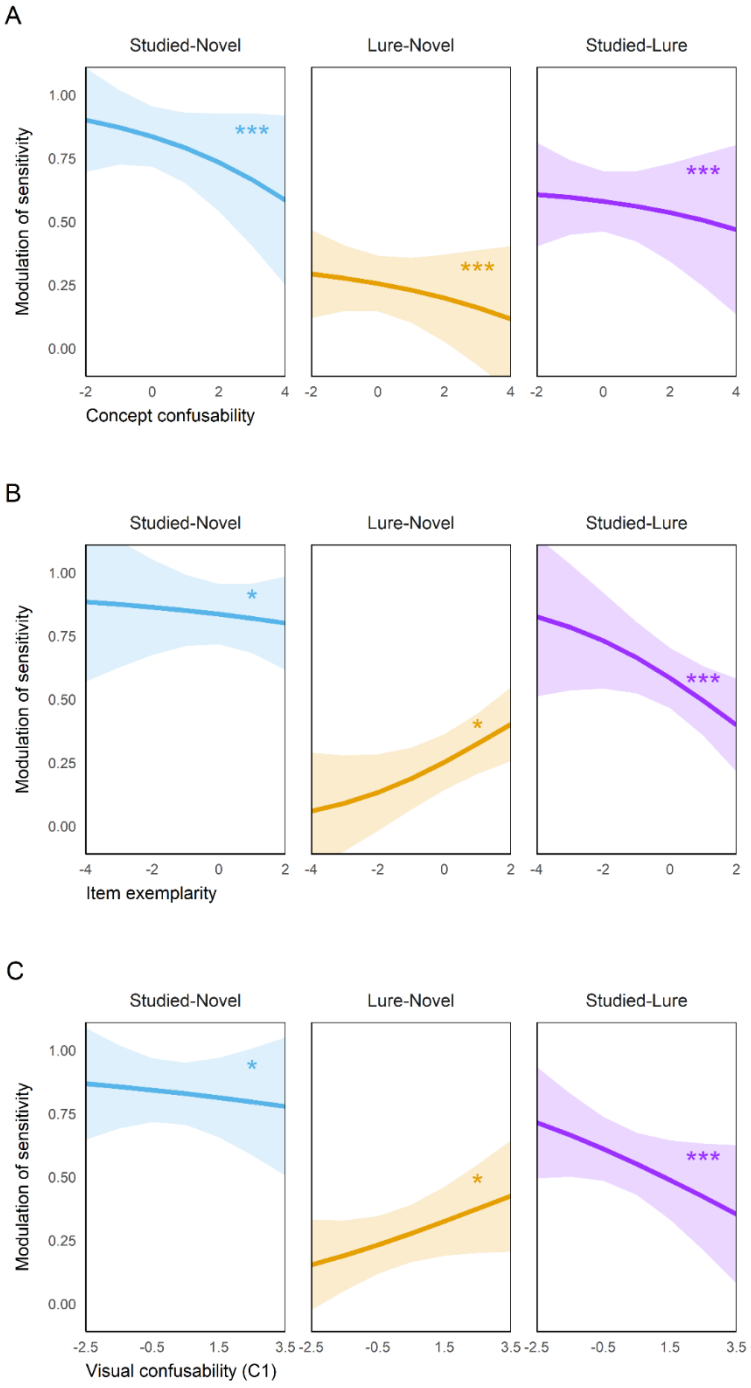
Results of Experiment 1 (Novel Items as Baseline)

Variable	Estimate	*d'*	SE	z-value	*p*
(Intercept)	-2.41	-1.37	0.12	-20.02	<.001
Lure	1.71	0.95	0.07	23.09	<.001
Studied	4.85	2.74	0.09	52.00	<.001
Number of Features	0.13	0.07	0.08	1.66	.136
Concept Confusability	0.28	0.15	0.08	3.44	.002
Visual Confusability (C1)	0.06	0.03	0.07	0.85	.438
Visual Confusability (C2)	0.20	0.10	0.09	2.28	.040
Color Confusability	0.04	0.02	0.08	0.47	.635
Item Exemplarity	0.18	0.09	0.07	2.55	.025
Lure × Number of Features	-0.31	-0.17	0.07	-4.25	<.001
Studied × Number of Features	-0.10	-0.05	0.09	-1.11	.332
Lure × Concept Confusability	-0.28	-0.15	0.07	-3.96	<.001
Studied × Concept Confusability	-0.54	-0.29	0.09	-6.30	<.001
Lure × Visual Confusability (C1)	0.16	0.10	0.07	2.28	.040
Studied × Visual Confusability (C1)	-0.19	-0.10	0.08	-2.24	.041
Lure × Visual Confusability (C2)	-0.15	-0.06	0.09	-1.67	.136
Studied × Visual Confusability (C2)	-0.11	-0.05	0.10	-1.07	.332
Lure × Color Confusability	-0.04	-0.02	0.08	-0.48	.635
Studied × Color Confusability	-0.11	-0.06	0.09	-1.17	.317
Lure × Item Exemplarity	0.20	0.13	0.08	2.57	.025
Studied × Item Exemplarity	-0.21	-0.10	0.09	-2.29	.040

Note. The reference level of condition is set to "novel". Parameter estimates (logOR), *d'* equivalent, standard errors, z-values, and FDR-corrected *p*-values are listed for condition, concept-level, and item-level variables in the winning (full) linear mixed model selected with AIC. Nearest neighbor perceptual confusability measures were reported in the model above. See Material and Methods, Variables of Interest, and Results for details. SE = Standard Error.

Figure 2

Effects of Semantic and Perceptual Variables on Mnemonic Sensitivity in Experiment 1



Note. Plots show effects of semantic and perceptual variables on modulation of sensitivity. The plot lines represent the effect of the predictor variables on the probabilities of endorsing studied items as "old" relative to novel items (light blue), lures as "old" relative to novel items (orange), and studied items as "old" relative to lures (purple). Panel A, B, and C show the effects of concept confusability, item exemplarity, and C1 visual confusability in Experiment 1. See Material and Methods, Variables of Interest, and Results for details. * $p < .05$; ** $p < .01$; *** $p < .001$ (FDR-corrected).

Sensitivity for Studied Relative to Novel Items

The probability of judging a studied item as "old" was compared to the probability of judging a novel item as "old". Overall, discrimination of studied from novel items was very good (simple effect on *d'* = 2.74; 95% CI: 2.65, 2.83). The results also showed modulations of concept- and item-level variables on sensitivity for studied items. Images whose concepts were more confusable with other concepts in the set were less likely to be remembered (Figure 2A; interaction of concept confusability with studied items on *d'* = -0.29; 95% CI: -0.38, -0.20). Pictures judged with high exemplarity were also less likely to be remembered (Figure 2B; interaction of item exemplarity with studied items on *d'* = -0.10; 95% CI: -0.19, -0.01). Lastly, participants were less likely to recognize items that were visually confusable in terms of their low-level representations (Figure 2C; interaction of C1 with studied items on *d'* = -0.10; 95% CI: -0.19, -0.01). No other variable significantly modulated sensitivity for studied items.

Sensitivity for Lure Relative to Novel Items

The probability of judging a lure item as "old" was compared to the probability of judging a novel item as "old". As expected, participants were more likely overall to incorrectly endorse lures than novel items (simple effect on *d'* = 0.95; 95% CI: 0.88, 1.03). The concept-level semantic variables also had substantial effects on lure sensitivity. Fewer errors were observed for lure images whose concepts were more confusable with other concepts in the set (Figure 2A; interaction of concept confusability with lure on *d'* = -0.15; 95% CI: -0.23, -0.07), as well as for those with a greater number of semantic features (interaction of number of features with lure on *d'* = -0.17; 95% CI: -0.25, -0.10). At the item-level, false recognition was more likely for lures rated as better exemplars of their concept (Figure 2B; interaction of item exemplarity with lure on *d'* = 0.13; 95% CI: 0.05, 0.21), as well as for lures whose early visual representations (such as line orientation) were more confusable (Figure 2C; interaction of C1 with lure on *d'* = 0.10; 95% CI: 0.03, 0.18). No other variable significantly modulated sensitivity for lure items.

Sensitivity for Studied Relative to Lure Items

The results of the analysis that examined the net modulation of participants' ability to discriminate studied items from lures are shown below in Table 2, and illustrated in Figure 2. Overall, performance was fairly good (simple effect on *d'* = 1.79; 95% CI: 1.71, 1.87). Both concept- and-item level variables modulated this effect. Mnemonic discrimination was poorer for items with high concept confusability (Figure 2A; interaction of concept confusability with studied items on *d'* = -0.13; 95% CI: -0.21, -0.06), but was better for concepts with a larger overall number of semantic features (interaction of number of features with studied items on *d'* = 0.12; 95% CI: 0.04, 0.20). At the item level, objects judged as having high item exemplarity were also less well discriminated (Figure 2B; interaction of item exemplarity with studied items on *d'* = -0.23; 95% CI: -0.31, -0.15), as were images with greater low-level (C1) visual confusability (Figure 2C; interaction of C1 with studied items on *d'* = -0.20; 95% CI: -0.28, -0.13). No other variable significantly modulated sensitivity for studied relative to lure items.

Table 2

Results of Experiment 1 (Lure Items as Baseline)

Variable	Estimate	*d'*	SE	z-value	*p*
(Intercept)	-0.70	-0.42	0.11	-6.39	<.001
Studied	3.14	1.79	0.07	42.02	<.001
Studied × Number of Features	0.21	0.12	0.08	2.80	.005
Studied × Concept Confusability	-0.25	-0.13	0.07	-3.49	<.001
Studied × Visual Confusability (C1)	-0.35	-0.20	0.07	-4.76	<.001
Studied × Visual Confusability (C2)	0.03	0.02	0.08	0.40	.686
Studied × Color Confusability	-0.07	-0.04	0.08	-0.93	.354
Studied × Item Exemplarity	-0.40	-0.23	0.08	-5.07	<.001

Note. The reference level of condition is set to "lure". Parameter estimates (logOR), *d'* equivalent, standard errors, z-values, and FDR-corrected *p*-values are listed for condition, concept-level, and item-level variables in the winning (full) linear mixed model selected with AIC. Nearest neighbor perceptual confusability measures were reported in the model above. See Material and Methods, Variables of Interest, and Results for details. SE = Standard Error.

False Alarms to Novel Items

Overall, participants were good at identifying unstudied novel items as "new" (intercept on *d'* = -1.37; 95% CI: -1.50, -1.25). This baseline was also modulated by both concept- and item-level semantic variables. False alarms were more frequent for highly confusable novel items (Supplemental Figure 5A; simple effect of concept confusability on *d'* = 0.15; 95% CI: 0.07, 0.23), and for those with high rated item exemplarity (Supplemental Figure 5B; simple effect of item exemplarity on *d'* = 0.09; 95% CI: 0.02, 0.17). The baseline probability of false alarms to novel items was also modulated by the C2 index of visual confusability. People were more likely to falsely endorse novel items as "old" if they were highly confusable in terms of late visual representations (such as global shape) (simple effect of C2 on *d'* = 0.10; 95% CI: 0.01, 0.19). No other variable significantly modulated baseline false recognition of novel items.

Discussion

In Experiment 1 we investigated the effects of conceptual and perceptual confusability on mnemonic discrimination of objects, using objective measures of similarity within a generalized linear mixed model framework. The findings for both concept-level and item-level variables support the proposal that semantic and perceptual properties shared by studied and lure objects contribute to mnemonic discrimination. At the concept level, studied objects whose concepts were more confusable because they shared features with other concepts were less likely to be remembered, and the corresponding lures were more likely to be correctly rejected. In contrast, at the item level, while studied objects that were better exemplars of their concept were less well recognized, lure objects with this property triggered more frequent errors. Likewise, people also misrecognized more lures that were visually confusable with their nearest neighbor. Overall, these effects resulted in poorer mnemonic discrimination between

studied and lure objects with high concept confusability, high item exemplarity, and high visual confusability.

The effects of concept confusability on sensitivity for studied items and lures suggest that when concepts were more confusable with other concepts, memory was weaker for the basic-level information shared by the studied items and lures. Our metric of concept confusability was based on an established model of conceptual structure and existing norms for feature properties of common concepts (Devereux et al., 2014; Tyler & Moss, 2001). This measure reflects conceptual processing of shared information across a set of items. The more confusable concepts shared more features with other concepts, and had fewer distinctive features not shared with other concepts. Our results suggest that if people remember less distinctive information about an object's concept, they are less likely to remember the object. The same logic applies to lures: lures with more confusable concepts can be more easily rejected as unstudied because memory is weaker for the concept shared with the studied item. This interpretation is supported by Taylor et al.'s (2012) finding that domain-level categorization decisions (living/ non-living judgments) were faster for concepts with more shared features, while basic-level naming was faster for concepts with more distinctive features. Thus, the efficiency of basic-level processing depended on the relative emphasis on coarse, cross-concept processing of shared features relative to fine-grained processing of individual concepts with distinctive features. Here, we did not find any modulation of study phase RT by concept confusability or other experimental variables. This probably reflected the use of self-paced judgments at study and prioritization of accuracy over speed given our primary concern with memory accuracy. From a fuzzy trace theory viewpoint, gist memory contributes positively to both veridical memory and lure false recognition (Brainerd et al., 2008; Brainerd et al., 1995). In this typical categorized pictures task, the gist is at the basic level since studied and lure objects are exemplars of the same basic-level concepts (e.g. a dog was studied, and a different dog appeared as a lure). In Experiment 1, this gist was impoverished for highly confusable concepts, reducing both recognition of the studied items and gist-related errors for the lures. Overall, mnemonic discrimination between these items was poorer.

While people were better able to successfully reject lures with more confusable concepts, false alarms to novel items with more confusable concepts were increased (see Supplemental Figure 5A). These items did not share a basic-level concept with any studied items but did share semantic features with them to varying degrees. This result is similar to that reported by Montefinese et al. (2015) for verbal stimuli. Participants studied sets of categorically related words (e.g., car, truck, scooter for the vehicle category) and were more likely to falsely recognize unstudied words which had more shared semantic features (e.g., tram). In that study, the novel words were related to the studied sets by membership of common superordinate categories, like the novel objects in the current study. Thus, it is possible that this higher-level semantic similarity caused these items to be more difficult to discriminate in memory. However, since both this result and Montefinese et al.'s (2015) finding were modulations of raw false alarms, we cannot rule out the alternative possibility that they reflect an effect of concept confusability on response criterion (Heit et al., 2003; Loiotile & Courtney, 2015).

The results for item exemplarity show that sensitivity for both studied items and lures depended on semantic processing at the item level as well as the concept level. As predicted, higher exemplarity lures were more likely to be falsely identified as studied, consistent with enhanced activation of basic-level representations. This suggests that participants tended to remember having studied the corresponding basic-level concept (e.g., giraffe in Figure 1) even if they did not remember the specific studied exemplar. These data are consistent with Barry et al.'s (1997) finding of facilitated basic-level naming for high exemplarity objects. However, studied items rated as better exemplars of their basic-level concept were more likely to be forgotten. This suggests that better exemplars were more likely to trigger reactivation of basic-level gist information at test, and perhaps less likely to elicit retrieval of specific information. It is not consistent with a simple modulation of basic-level gist at encoding, which should increase true recognition as well as lure errors.

Our initial analysis of the effects of perceptual relations on memory used graded indices of confusability computed across the whole set of stimuli (see Supplemental Figure 4 and Table 7). However, since the graded weighting prioritized similarity over many items, the high confusability items had generic visual properties rather than being similar to any specific items in the set. These observations motivated the analysis using perceptual confusability measures with a stronger weight, restricted to nearest neighbors only. As Figure 1 shows, these metrics had a quite different profile from the across-set measures (compare to Supplemental Figure 4). For C1, more confusable items were those with a distinctive linear orientation, shared with at least one other specific item in the set (the neighbor), although more color-confusable items were relatively uniform as well as bland. The analysis using these measures revealed that lures that shared low-level (C1) visual features with their nearest neighbor were more likely to be misrecognized. However, the corresponding studied items were more likely to be forgotten. We consider these effects of nearest neighbor visual confusability further below, in light of the results of Experiment 2.

The data point to multiple gist-like effects on mnemonic discrimination reflecting both semantic and perceptual dimensions along which studied items and lures were similar to other studied objects. Here, the lures were different exemplars of studied basic-level concepts, so variables increasing emphasis on processing at this level tended to trigger errors, while variables indexing processing shared with other concepts tended to reduce them. We have proposed that these effects reflect strengthening or weakening of basic level conceptual memory representations. However, there is an alternative mechanism by which lures with more confusable concepts could be better discriminated: they might be more easily rejected using the memory editing strategy of recall-to-reject (Gallo, 2004; Brainerd et al., 2003). When a lure triggers recollection of similar studied items, people can avoid gist-related errors by comparing the recollected information with the lure (e.g., they decide that a white dog was not presented, because they remember having studied a black dog). A similar mechanism might also apply to differences in performance for lures that were related to studied items in other ways, at the item level. It was important to establish whether our results would generalize to a situation in which recall-to-reject was prevented.

Experiment 2

In Experiment 2 we aimed to replicate Experiment 1's novel semantic and perceptual effects on memory, and test whether these modulations of mnemonic discrimination would generalize to a task where the use of a recall-to-reject strategy was prevented. We therefore amended the procedure so that multiple different exemplars of each basic-level concept were studied. Once people have to recall more than 5 different studied items in order to reject a single lure, a recall-to-reject strategy becomes ineffective, particularly if the set size varies within the study list (Gallo, 2004). In Experiment 2, participants studied sets of either two or eight different exemplars of the same basic-level concept. Our preregistered prediction was that if Experiment 1's finding of reduced false recognition of lures with more confusable concepts was due to impoverished representations of these basic-level concepts in memory, Experiment 2 would show a similar effect. With more studied exemplars per concept we might also observe enhanced effects of concept confusability on mnemonic discrimination in Experiment 2 compared to Experiment 1, and for set size 8 compared to set size 2. However, if the recall-to-reject account is correct, concept confusability would either be associated with *increased* lure errors in Experiment 2, or would have no effect. In terms of item-level processing, we expected that effects might be similar, or more pronounced because of the larger studied sets.

Materials and Methods

The experimental methods were preregistered with the Open Science Framework (https://osf.io/3h7kf).

Participants

The study included sixty adults aged 18-33 years (M = 21.2; SD = 3.2, 12 male, 48 female). A further ten participants were excluded from data analysis: 9 due to technical issues with recording responses, 1 due to poor performance at test (using the preregistered criterion of *d'* for studied item discrimination of less than 3 SD from the mean). The sample size was determined with sensitivity analyses setting alpha to .05. The principal effect of interest was the overall effect of concept confusability on lure false recognition (interaction of concept confusability × lure versus novel baseline; collapsed across set size). Simulations with simR suggested that with N = 60 we would have .92 power to detect a small effect (OR = 0.70; equivalent to Cohen's d = 0.20; the effect size in Experiment 1 was equivalent to d = 0.15, although as outlined above, we expected this effect to be larger, if present). For the higher order interaction of concept confusability with study set size condition (lures for set size 8 versus 2), N = 60 could detect a medium-sized interaction of concept confusability × condition × lure versus novel baseline at 0.99 power (OR = 0.58; equivalent to Cohen's d = 0.30). Data from a group of older participants collected at the same time will be included in a separate report. All participants were fluent English speakers (since at least the age of 5), and had normal or corrected to-normal vision. Participants were recruited by local advertisement and provided informed consent. They received either course credit or an honorarium. The study was approved by the University of Edinburgh Psychology Research Ethics Committee (Ref. 278-1617/1).

Stimuli

Except where specified, stimuli were the same as in Experiment 1. For each of the 200 basic-level concepts, 9 different sets of images were obtained for a total of 1800 images. Each study list included 600 items: 480 in the large sized sets (i.e., set size 8) and 120 in the small sized sets (i.e., set size 2). Each test list consisted of 300 items: 120 studied images (60 from set size 8 and 60 from set size 2), 120 similar lures (60 unstudied exemplars of studied basic-level concepts from set size 8 and 60 from set size 2, and 60 novel items whose basic-level concepts had not been studied). Three filler trials prefaced both the study and the test phase. We generated one study and test list for each participant which randomized the allocation of the concepts and their exemplar images to conditions (i.e., studied, lure, and novel) and set size (2 and 8) with the constraint that half the concepts in each condition (item type and set size) were living and half non-living.

Procedure

The experiment consisted of a single study phase followed by a recognition test phase with interspersed standardized cognitive tests (not reported here) for 15 minutes. Stimuli were presented with MATLAB (R2018b, The MathWorks) using PsychToolbox (Kleiner et al., 2007; Version 2.0.14). The procedure was otherwise the same of Experiment 1 except that trials during the study phase were not self-paced, but presented every 3 s.

Statistical Analysis

The variables of interest and nuisance variables (see Supplemental Table 8) were identical to those used in Experiment 1, but the item metrics were recomputed for this larger set of images. Analyses reported here were based on the nearest neighbor metrics of perceptual confusability for reasons noted in Experiment 1, Materials and Methods, Variables of Interest, and the results for graded perceptual confusability metrics are given in the Supplemental Material (Supplemental Table 11). Item exemplarity ratings were again collected from the participants after the test phase. As the concepts were identical to those used in Experiment 1, we used the same concept familiarity values collected previously. The main preregistered statistical analyses collapsed across the large and small study set sizes were identical to those used in Experiment 1, as were the model selection procedures. We also examined *a priori* the modulatory effects of set size (i.e., set size 8 vs set size 2), but as this variable had no significant effects we focus here on the results collapsed over the two set sizes (see Supplemental Table 10 for the results of the full model). Lastly, we again conducted a linear mixed-effect analysis of study phase RTs (results not reported, see https://osf.io/ndk83/).

Results

As in Experiment 1, there was decisive evidence in favor of the full model including concept-level, item-level, and confounds principal component variables, relative to simpler models (see Supplemental Table 9). The full model was the most parsimonious, receiving the lowest AICc score (AICc = 17597.96), and substantial weight (AICc weight = 0.89) relative to the other models, with an evidence ratio between the top-ranked model and the second-ranked model of 7.83. This result was also supported by a significant LRT in favor of the full relative to the model ranked second on the basis of AIC ($\chi2(7) = 18.16$, $p = 0.011$). Unlike Experiment 1, BIC

now provided evidence in favor of the model with both concept- and-item level variables as the most parsimonious model (BIC = 17780.99). The results of the mixed effects analysis of the full model are shown below in Table 3, and illustrated in Figure 3. All the *p*-values are FDR-multiple comparison corrected (Benjamini & Hochberg, 1995). Raw recognition responses by item type are reported in Supplemental Figure 6.

Table 3

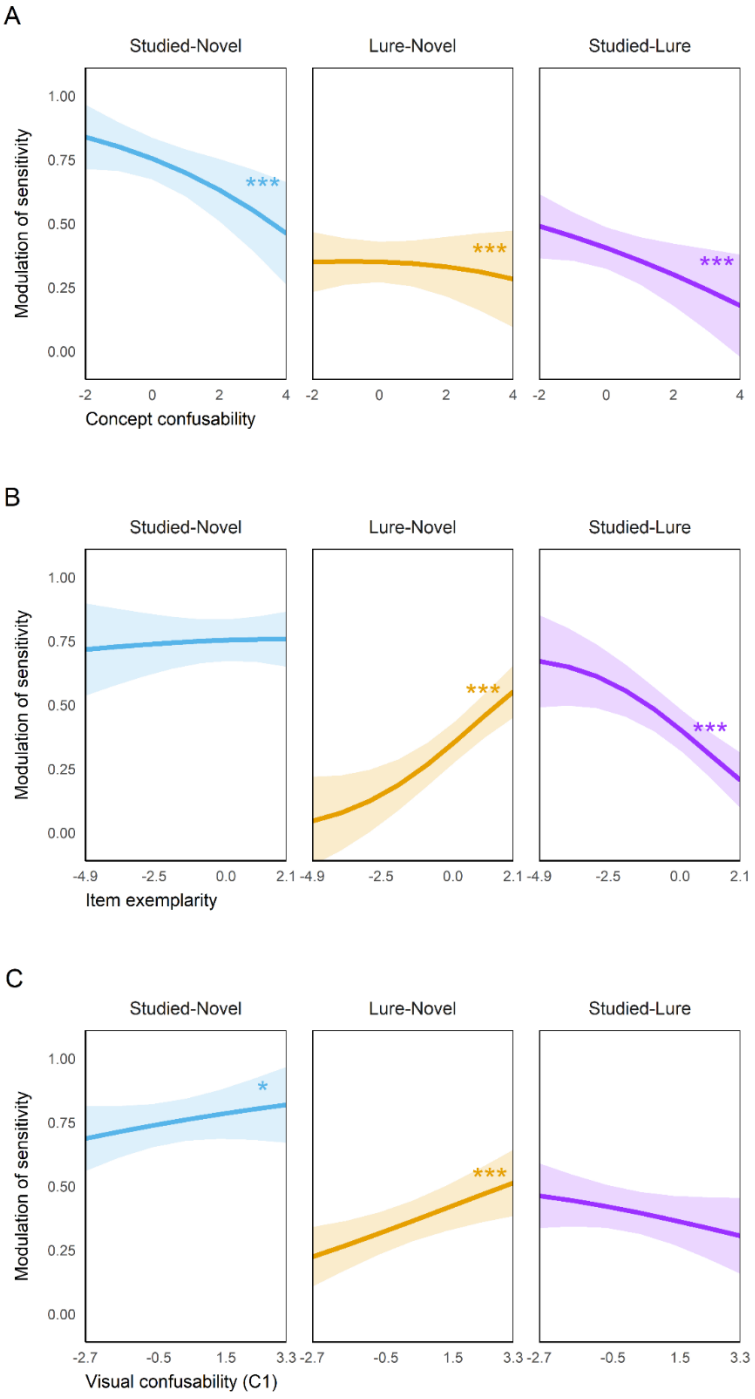
Results of Experiment 2 (Novel Items as Baseline)

Variable	Estimate	*d'*	SE	z-value	*p*
(Intercept)	-2.70	-1.55	0.10	-26.65	<.001
Lure	2.34	1.33	0.07	32.00	<.001
Studied	4.20	2.44	0.08	54.42	<.001
Number of Features	0.00	0.00	0.07	-0.03	.978
Concept Confusability	0.31	0.16	0.07	4.30	<.001
Visual Confusability (C1)	-0.08	-0.03	0.07	-1.13	.364
Visual Confusability (C2)	0.22	0.10	0.08	2.83	.011
Color Confusability	0.07	0.03	0.07	1.07	.375
Item Exemplarity	0.14	0.08	0.07	2.11	.067
Lure × Number of Features	-0.01	0.00	0.07	-0.12	.949
Studied × Number of Features	-0.09	-0.05	0.07	-1.28	.301
Lure × Concept Confusability	-0.25	-0.12	0.07	-3.59	<.001
Studied × Concept Confusability	-0.53	-0.29	0.07	-7.32	<.001
Lure × Visual Confusability (C1)	0.26	0.14	0.07	3.66	<.001
Studied × Visual Confusability (C1)	0.20	0.10	0.07	2.66	.016
Lure × Visual Confusability (C2)	-0.16	-0.06	0.08	-1.95	.082
Studied × Visual Confusability (C2)	-0.17	-0.07	0.08	-2.05	.071
Lure × Color Confusability	0.05	0.04	0.07	0.70	.536
Studied × Color Confusability	-0.07	-0.03	0.07	-0.95	.422
Lure × Item Exemplarity	0.29	0.18	0.07	4.03	<.001
Studied × Item Exemplarity	-0.06	-0.03	0.07	-0.82	.483

Note. The reference level of condition is set to "novel". Parameter estimates (logOR), *d'* equivalent, standard errors, *z*-values, and FDR-corrected *p*-values are listed for condition, concept-level, and item-level variables in the winning (full) linear mixed model selected with AIC. Nearest neighbor perceptual confusability measures were included in the model above. See Material and Methods, Variables of Interest, and Results for details. SE = Standard Error.

Figure 3

Effects of Semantic and Perceptual Variables on Mnemonic Sensitivity in Experiment 2



Note. Plots show effects of semantic and perceptual variables on modulation of sensitivity. The plot lines represent the effect of the predictor variables on the probabilities of endorsing studied items as "old" relative to novel items (light blue), lures as "old" relative to novel items (orange), and studied items as "old" relative to lures (purple). Panel A, B, and C show the effects of concept confusability, item exemplarity, and C1 visual confusability in Experiment 2. See Material and Methods, Variables of Interest, and Results for details. * $p < .05$; ** $p < .01$; *** $p < .001$ (FDR-corrected).

Sensitivity for Studied Relative to Novel Items

As in Experiment 1, discrimination of studied from novel items was very good (simple effect on *d'* = 2.44; 95% CI: 2.36, 2.51). Concept confusability impaired sensitivity for studied items which were more likely to be forgotten (Figure 3A; interaction of concept confusability with studied items on *d'* = -0.29; 95% CI: -0.36, -0.21). However, the effect of item exemplarity on true recognition was no longer significant (Figure 3B; interaction of item exemplarity with studied items on *d'* = -0.03; 95% CI: -0.10, 0.05). Also, more visually similar studied items were now *more* (rather than less) likely to be correctly recognized (Figure 3C; interaction of C1 with studied items on *d'* = 0.10; 95% CI: 0.03, 0.18). No other variable significantly modulated sensitivity of studied items.

Sensitivity for Lure Relative to Novel Items

As in Experiment 1, participants were generally more likely to judge lures than novel items as "old" (simple effect on *d'* = 1.33; 95% CI: 1.26, 1.41). The direction of the concept confusability effect was also unchanged: false recognition of lures was again relatively *less* likely for pictures whose concepts shared many semantic features with other concepts (Figure 3A; interaction of concept confusability with lure on *d'* = -0.12; 95% CI: -0.20, -0.05). At the item level, lure errors were again more frequent for items with higher rated exemplarity (Figure 3B; interaction of item exemplarity with lure on *d'* = 0.18; 95% CI: 0.11, 0.26), and with higher visual confusability in terms of low-level visual representations (Figure 3C; interaction of C1 with lure on *d'* = 0.14; 95% CI: 0.07, 0.22). No other variable significantly modulated sensitivity for lure items.

Sensitivity for Studied Relative to Lure Items

The results of the analysis that examined participants' ability to discriminate studied items from lures are shown below in Table 4, and illustrated in Figure 3. Overall, participants were fairly good at discriminating studied items from similar lures (simple effect on *d'* = 1.10; 95% CI: 1.06, 1.15). Both concept- and-item level variables modulated this effect. Similar to Experiment 1, studied items with high concept confusability were less likely to be correctly discriminated from highly confusable lures (Figure 3A; interaction of concept confusability with studied items on *d'* = -0.16; 95% CI: -0.21, -0.12). However, unlike Experiment 1, more semantic features did not improve discrimination (interaction of number of features with studied items on *d'* = -0.05; 95% CI: -0.10, 0.00). Studied concepts whose pictures were judged to have high exemplarity were again less likely to be discriminated from their lure exemplars (Figure 3B; interaction of item exemplarity with studied items on *d'* = -0.21; 95% CI: -0.26, -0.16). Lastly, participants were less likely to correctly discriminate studied items with high color confusability (interaction of color confusability with studied items on *d'* = -0.07; 95% CI: -0.12, -0.02). Unlike Experiment 1, the effect of C1 visual confusability was no longer significant (Figure 3C; interaction of C1 with studied items on *d'* = -0.04; 95% CI: -0.09, 0.01). No other variable significantly modulated sensitivity for studied relative to lure items.

Table 4

Results of Experiment 2 (Lures as Baseline)

Variable	Estimate	*d'*	SE	z-value	*p*
(Intercept)	-0.35	-0.21	0.08	-4.45	<.001
Studied	1.85	1.10	0.04	44.54	<.001
Studied × Number of Features	-0.09	-0.05	0.04	-1.91	.079
Studied × Concept Confusability	-0.28	-0.16	0.04	-6.65	<.001
Studied × Visual Confusability (C1)	-0.06	-0.04	0.04	-1.47	.186
Studied × Visual Confusability (C2)	-0.01	-0.01	0.04	-0.25	.847
Studied × Color Confusability	-0.12	-0.07	0.04	-2.89	.007
Studied × Item Exemplarity	-0.35	-0.21	0.04	-8.32	<.001

Note. The reference level of condition is set to "lure". Parameter estimates (logOR), *d'* equivalent, standard errors, z-values, and FDR-corrected *p*-values are listed for condition, concept-level, and item-level variables in the winning (full) linear mixed model selected with AIC. Nearest neighbor perceptual confusability measures were reported in the model above. See Material and Methods, Variables of Interest, and Results for details. SE = Standard Error.

False Alarms to Novel Items

Participants correctly identified most unstudied novel items as "new" (intercept on *d'* = -1.55; 95% CI: -1.65, -1.44). As predicted, and as in Experiment 1, items with high concept confusability were more likely to be falsely recognized as "old" (Supplemental Figure 6A; simple effect of concept confusability on *d'* = 0.16; 95% CI: 0.09, 0.24). At the item level, false alarms to novel items were modulated by C2 visual confusability. People were more likely to misrecognize novel items with more confusable late visual representations (i.e., their overall shape) (simple effect of C2 on *d'* = 0.10; 95% CI: 0.02, 0.17). No other variable significantly modulated baseline false recognition of novel items.

Discussion

Experiment 2 closely reproduced Experiment 1's procedures, except that participants studied multiple exemplar images of each basic-level concept. This allowed us to test whether the variables contributing to mnemonic discrimination were altered when people could not effectively use a recall-to-reject strategy. The effect observed in Experiment 1 for concept confusability was qualitatively the same in Experiment 2: people were again less likely to misrecognize lures with high concept confusability (reductions in *d'* by factors of -0.15 and -0.12 in Experiments 1 and 2), and more likely to forget more confusable studied items (*d'* reductions of -0.29 and -0.29 in Experiments 1 and 2). Therefore, the pattern of findings in Experiment 1 cannot be explained by a facilitation of recall-to-reject for lures with more confusable concepts. Instead, the data suggest that processing shared features over concepts weakened the representation of basic-level conceptual information in memory. Thus, concept confusability reduced recognition of both studied and lure exemplars and impeded the ability to discriminate between them in memory.

As in Experiment 1, semantic relations over items also had a marked effect on lure errors, which were again more frequent for lures that were better exemplars of their concept (*d'* increasing by factors of 0.13 and 0.18 in Experiments 1 and 2). This suggests that at test, people were more likely to respond "old" to high exemplarity lures because they remembered having studied the corresponding basic-level concepts. However here, unlike in Experiment 1, item exemplarity did not significantly affect true recognition of studied items (*d'* effects of -0.10 and -0.03 in Experiments 1 and 2). We do not place much emphasis on this null finding, since the between-Experiment interaction between study set size, item exemplarity and old versus novel items was also not significant (see Supplemental Table 12). It may be that there is an effect on true recognition which is too small for us to detect consistently. Alternatively, the effect observed in Experiment 1 may have been weakened by the increase in set size: for example, exemplarity of individual studied items might matter less when multiple exemplars are studied. Despite this, the effect for lures remained robust, as did the effect on overall discrimination between studied items and lures.

For the perceptual item-level variables, as already noted in Experiment 1, the graded measures did not seem to capture our initial intuition that some lures would be highly visually confusable with specific studied items. Therefore, we focused instead on exploratory analyses that better tested our original prediction that such lures would be more frequently misrecognized. The nearest neighbor metrics yielded consistent findings in the two experiments in this regard. For the low-level C1 measure, lures that were more visually confusable with another item were more likely to be misrecognized (*d'* increased by factors of 0.10 and 0.14 for Experiments 1 and 2). However, the effects of C1 visual confusability on recognition of studied items differed between the two experiments. While in Experiment 1 more confusable items were less likely to be remembered, in Experiment 2 they were *more* often remembered (*d'* modulations of -0.10 and 0.10). This reversal may reflect a genuine difference due to the increase in study set size in Experiment 2: for example, if more highly similar nearest neighbors were introduced by use of multiple studied exemplars of each basic-level concept. Further data will be required to establish whether this is a robust finding. As a result of this different effect for studied items, visual confusability did not significantly reduce overall mnemonic discrimination of studied items from lures in Experiment 2, unlike for Experiment 1. The current results point to a particular salience of simple visual features like line orientation for false recognition of lures (see Figure 1). We explore the possible reasons for this below.

General Discussion

In this research, we used objective and model-based measures to show for the first time that multiple semantic and perceptual relations contribute to people's ability to discriminate objects in memory. This approach allowed us to assess simultaneous influences of semantic and perceptual relations between objects while controlling for the potential effects of other variables known to influence mnemonic discrimination. Using generalized linear mixed model analysis we were also able to directly model the effects of predictors on binary memory outcomes, and generalize the results over concepts as well as participants (DeCarlo, 1998; Wright et al., 2009). In both experiments, studied objects that shared semantic features with

many other concepts were more difficult to recognize, and lures whose corresponding studied object had many shared features were easier to correctly identify as new. In contrast, misrecognition was more frequent for lure objects that were more representative of their basic-level concept. The findings demonstrate distinct effects of semantic relations among concepts, and semantic relations between concepts and their exemplars. Simple perceptual properties shared with other studied images also contributed to misrecognition of lures, suggesting that image as well as concept properties are important in mnemonic discrimination for pictures.

To assess the effects of semantic similarity on memory we used a concept confusability metric derived from a feature-based model of semantic memory (Devereux et al., 2014; Tyler & Moss, 2001). Previous studies have shown that conceptual structure understood in terms of feature relations between concepts can explain a range of phenomena, such as differences in the processing of living and nonliving concepts in healthy people (McRae et al., 1997; Randall et al., 2004; Taylor et al., 2011; Vigliocco et al., 2004) and specific impairments in neuropsychological patients (Forde & Humphreys, 1999; Humphreys & Forde, 2001; Warrington & Shallice, 1984). For example, Moss et al. (1997) described a post-encephalitic patient who was very poor at differentiating between highly similar objects (e.g., tiger versus panther), but had no difficulty in determining the superordinate category of an object (e.g., land animals; see also Tyler et al., 2004). This dissociation between finer-grained and coarser categorical levels of conceptual processing converges with Taylor et al.'s (2012) finding that, in healthy people, processing highly shared semantic features facilitated domain-level categorization decisions but impeded basic-level naming.

The current study is the first to show that feature-based conceptual structure impacts mnemonic discrimination. The results converge with earlier evidence from Montefinese et al. (2015) that semantic feature similarity increases false alarms to unstudied items (see also Montefinese et al., 2018). Our data confirm that overlapping semantic features impact mnemonic sensitivity indices that adjust false recognition for response criterion, and further show that concept confusability also affects true recognition. The results are also in line with Coane et al. (2016)'s suggestion that specific effects of semantic categorical relations on false recognition reflect feature similarity at least in part (see also Coane et al., 2020). We have interpreted this finding in terms of gist memory traces representing basic-level conceptual information. According to fuzzy trace theory, memory outcomes reflect the relative accessibility of two kinds of memory traces encoded in parallel: verbatim traces containing specific representations of the studied items, and gist traces representing their meaning traces (Brainerd & Reyna, 2002). Memory for studied items is supported by both verbatim and gist traces, and false memory for related lures occurs when gist but not verbatim traces are retrieved. Although the theory does not specify the informational content of gist, the logic of the mnemonic discrimination task suggests that the relevant gist information in any given task is the information that is shared between the studied items and lures. Our results support this assumption, and suggest that different levels of semantic similarity have distinct effects on mnemonic discrimination. In the categorized pictures task used here, studied items and lures shared basic-level concepts. Concept confusability – a similarity metric reflecting shared processing of conceptual features – weakened the representations of individual concepts in memory and therefore reduced gist-like

effects on memory for studied items and lures. The net effect on mnemonic discrimination was also negative: for concepts that were more confusable, people were less able to discriminate in memory between studied items and lures.

In contrast, item exemplarity – similarity between an exemplar item and its concept – strengthened activation of the concept and therefore increased gist-like effects, at least for lures. Exemplarity, or image agreement, defines a relation between an image exemplar and its basic-level concept, and is closely related to typicality (Barry et al., 1997; Snodgrass & Vanderwart, 1980). A typical exemplar shares more features with other members of its category (Rosch & Mervis, 1975) and may therefore more strongly elicit a gist memory trace that overlaps with representations of other exemplars. In line with this, we found that people falsely recognized more high exemplarity basic-level lures. Together with the findings for concept confusability, the data suggest that both concept-level and item-level variables impacted mnemonic discrimination via modulations of basic-level memory representations. A few prior studies have examined the specific semantic information that gist is based on. These have used specific stimuli or measures that focused on a single type of relation (Cann et al., 2011; Coane et al., 2016; Montefinese et al., 2015). Our use of multiple measures enabled effects of relations at different levels to be examined at the same time.

As outlined in the Introduction, studies using a range of different task materials also suggest that the content of the memory representations critical for mnemonic discrimination varies. Studies using pictures, words, situational themes, and narratives as to-be-remembered material find that lures that are related to studied items at these different semantic levels are misrecognized as studied, implicating gist representations at different levels. The results for novel items offer converging evidence that the effects of different semantic relations on lure errors may depend on the nature of the lures to be discriminated. In both experiments, while concept confusability improved rejection of lures, it increased false alarms to unrelated novel images. Although neither these objects (e.g., a panther) nor other exemplars of their basic-level concepts had been studied (e.g., a different panther), people had studied other items from the same superordinate categories (e.g., a cat and a dog) as well as further items with which they shared semantic features (e.g. a chair, which < has legs >). Since no basic-level gist memory was ever encoded for these items' concepts, its effects could not be impoverished by emphasis on features shared with other concepts. However, if the coarse semantic activation not only reduced within-concept gist but enhanced across-concept gist, it would lead some novel objects with many shared features to be classified as "old". A similar argument was made by Montefinese et al. (2015) for verbal material. Our finding is also in line with previous studies showing that novel pictures (Bowman et al., 2019; Seamon et al., 2000) and words (Brainerd et al., 1995; Coane et al., 2016, 2020) belonging to the same superordinate category as studied items are more likely to be misrecognized than unrelated novel items. The current study and that of Montefinese et al. (2015) are the first to directly demonstrate that shared semantic features can increase false recognition of pictures and words, respectively. However, both these findings could be due to modulations of response criterion rather than effects on memory (see also Montefinese et al., 2018). People's bias to respond "old" versus "new" can change item-by-item in response to properties of the test probes (Heit et al., 2003; Kent et al., 2018). For

example, processing shared semantic features might increase attention to items in the absence of any retrieval of a gist trace. Further studies with an additional unrelated novel item baseline condition will be required to adjudicate between these two possibilities.

We have discussed the findings for both concept confusability and item exemplarity in terms of their effects on concept gist shared by studied items and lures. However, unlike concept confusability, item exemplarity did not affect memory for studied items as would be expected according to fuzzy trace theory if it modulated encoding of gist traces. Recognition of higher exemplarity studied items was numerically *reduced* in both experiments, significantly so in Experiment 1. The effects of exemplarity or typicality on recognition memory are relatively unexplored, but several studies suggest that words that are more typical members of their superordinate category tend to be better recalled, mediated in part by clustering by category (Schmidt, 1996). One possibility is that the restriction of the effect of exemplarity to lure errors derived from the way that exemplarity as an item-level variable assessed gist. Unlike concept confusability, which directly indexed shared features among concepts, item exemplarity indirectly indexed shared features among exemplars via judgments of the strength of the relations between exemplars and their concepts. Prioritization of relations between exemplars and concepts may have increased sensitivity to test phase cueing of memory traces by images eliciting strong basic-level representations. Greater overlap between an exemplar and its basic-level concept may therefore have increased the degree to which a lure image could trigger retrieval of a gist trace, but not the degree to which a studied image could elicit gist encoding at this level. Currently, norms are not available for the semantic features that are shared between individual pictured exemplars of a concept but, based on the assumptions of fuzzy trace theory, we would predict that similarity between studied items and lures on such measures would be associated with increases in recognition of both types of items, and reduced discrimination between the two.

The current data also support the proposal that mnemonic discrimination errors are driven by perceptual as well as semantic similarity (Bowman et al., 2019; Brady et al., 2008; Burnside et al., 2017; Koutstaal & Schacter, 1997; Motley & Kirwan, 2012; Seamon et al., 2000). We examined the effects of perceptual relations on memory using item-level measures of visual properties of the images. Lure items with greater low-level (i.e. C1) visual confusability with another image in the set were more likely to be misrecognized as "old". As these nearest neighbor analyses were exploratory, the results need to be treated with some caution, but the metrics yielded consistent results across the two experiments in the direction originally predicted for visual confusability effects. The findings suggest that perceptual effects on false recognition of lures are robust, at least for early visual properties like line orientation. The contrasting lack of consistent effects of perceptual confusability measured with graded metrics points to a dependence of this effect on specific similarity between lures and individual items in the set. We considered whether defining measures of perceptual (and semantic) similarity only between test items and each participant's set of studied items might provide a clearer picture; however, this was not possible due to very high collinearity with the original indices (Pearson's r between .80 and .98 in Experiments 1 and 2).

Our data converge with the handful of previous studies that have directly addressed perceptual effects. It is difficult to draw strong conclusions about color effects since we did not find consistent effects over experiments, and Reppa et al. (2020) (see Introduction) used a very different task and measures from ours. However, the current results are broadly in line with their finding that perceptual similarity (in their case, of shape) can cause interference in memory. Our data also converge with those of Brady et al. (2008) who found that people made errors to lures that were different-view images of studied objects, as well as to different exemplars of studied concepts. They used a forced choice recognition task, in which discrimination performance is generally substantially better than in old/new recognition, and may rely more on familiarity (Migo et al., 2009) and/or processing fluency (Voss et al., 2012). Our findings (and those of Motley and Kirwan, (2012); see Introduction) broadly converge with these earlier data to suggest that perceptual as well as semantic similarity both influence mnemonic discrimination regardless of the recognition task format. Our model-based measures specifically implicate low-level visual similarity in this task, but future studies may show that higher-level attributes like shape and view also contribute independently.

Others have found that perceptually similar but pre-experimentally meaningless lures trigger false recognition (Koutstaal et al., 2003; Pidgeon & Morcom, 2014; Slotnick & Schacter, 2004). For example, Slotnick and Schacter (2004) found that meaningless shapes that were visually related to studied items were more likely to be misrecognized relative to novel unrelated shapes. While visual similarity was not formally measured, lures tended to have similar linear orientation to their studied items. Our data are also in line with fMRI studies that have shown engagement of early visual cortex during false recognition of visually similar picture lures, whether or not these are also semantically related, supporting the interpretation that such activity represents enhanced visual processing during lure misrecognition (Bowman et al., 2019; Garoff-Eaton et al., 2006; Gutchess & Schacter, 2012; Slotnick & Schacter, 2004). Thus, behavioral and neuroimaging results converge on the idea that mnemonic discrimination errors can be elicited by both semantic and perceptual relations. The use of model-based metrics allowed us to go further by indexing perceptual properties directly, and addressing concerns that people may bring conceptual processing to bear even on experimentally unfamiliar stimuli (Pidgeon & Morcom, 2016).

These findings support earlier suggestions that gist-based memory can be perceptual as well as conceptual (Koutstaal & Schacter, 1997). This possibility is consistent with fuzzy trace theory's opponent processes, if gist is not restricted to information about meaning. A fuller understanding of mnemonic discrimination needs to take into account the nature of the relations and representations involved. Our finding that semantic relations between concepts and items had different effects on mnemonic discrimination suggests that different semantic relations may impact memory in ways that are modulated by task demands. An important question to address in future studies will be whether manipulations previously shown to impact measures of gist reliance (Brainerd & Reyna, 2005) can be shown to have dissociable effects on the contributions of semantic and perceptual similarity to mnemonic discrimination. Our results are also in line with previous data suggesting that semantic influences on memory errors go beyond associative activation (Brainerd et al., 2008; Cann et al., 2011; Coane et al., 2016, 2020)

The activation/monitoring theory explains false recognition in terms of associative strength (Roediger & McDermott, 2000; Roediger et al., 2001). On this view, studying a list of words produces an automatic associative activation that spreads through the lexical semantic system, and false recognition occur when lures accrue substantial activation through this process. A strength of the activation/monitoring theory is that the associative relations assumed to determine memory errors are quantifiable (Roediger et al., 2001). However, this account cannot explain the above findings. It is also inconsistent with perceptual effects on lure errors as it specifies that misrecognition stems from activation of lure representations at study. Similar lures which have no pre-existing associations to studied items are unlikely to be spontaneously generated at encoding (Arndt, 2010). Associative activation also cannot explain lure errors reflecting spatial proximity to the location of a studied object (Reagh et al., 2016, 2014).

Global-matching models can explain lure errors that reflect perceptual as well as semantic relations (Arndt & Hirshman, 1998; Arndt, 2010; Hintzman, 1988). According to models like MINERVA2 (Arndt & Hirshman, 1998), there is no gist memory trace, but mnemonic discrimination errors occur to the degree that lures presented at test are globally similar to multiple traces previously stored at encoding, without specification of the nature of the similar features. Such a retrieval-based mechanism may also explain why some variables – here, item exemplarity and visual confusability – impact lure errors but not memory for studied items. In fuzzy-trace theory's conjoint recognition model, lure-specific factors could modulate the probability of similarity responding at test, and therefore lure errors (Brainerd et al., 1999; Brainerd & Wright, 2005), although its processing tree model does not explicitly separate encoding and retrieval processes. An explicit model of encoding and retrieval contributions is offered by the pattern separation and completion processes at the heart of the complementary learning systems model (Marr, 1971; McClelland et al., 1995; McNaughton & Morris, 1987; Norman & O'Reilly, 2003). At study, pattern separation by the hippocampus ensures that specific memory traces are generated. When this separation fails memory traces may become more gist-like, and lures more likely to trigger errors (Wilson et al., 2006; Yassa & Stark, 2011). At test, people may also fail to discriminate lures because the lures trigger pattern completion due to high overlap with studied items (Motley & Kirwan, 2012; Norman & O'Reilly, 2003; Norman, 2010; Yotsumoto et al., 2007). Hippocampal pattern separation and completion do not specify the types of similarity more likely to influence mnemonic discrimination, and predict influences of perceptual, semantic and contextual properties (Hunsaker & Kesner, 2013; Reagh et al., 2014; Yassa & Stark, 2011). The complementary learning systems model further specifies non-pattern-separated neocortical inputs that may also contribute to semantically-driven mnemonic discrimination errors (Norman & O'Reilly, 2003; Pidgeon & Morcom, 2014; Wilson et al., 2006). These model predictions about specific encoding and retrieval operations are difficult to test with behavioral measures alone, but can be more directly investigated using neuroimaging measures, which enable semantic and perceptual modulations of neural processing during study and test phases to be assessed.

Conclusions

In this work we used objective measures derived from established models of conceptual structure and low-level vision to show that semantic and perceptual relations can simultaneously contribute to gist-like effects in memory. The results from two experiments implicated relations at multiple representational levels and suggested that similarity between studied and unstudied items does not always impair mnemonic discrimination. The coarse semantic activation elicited by processing shared semantic features across concepts impeded memory for studied objects while reducing false alarms to lure exemplars of the same concepts. In contrast, a strong semantic overlap at the item-level between an object and its basic-level concept was associated with more frequent false recognition of lures, as was strong low-level visual similarity between an object and a studied image. The initial findings were replicated in the second experiment which increased and varied the study set size in order to rule out recall-to-reject as an explanation for the results. Taken together, our findings point to the utility of a more structured and formal approach to understanding the relations underpinning gist-like effects in memory, and highlight the importance of image as well as concept properties in mnemonic discrimination.

References

Arndt, J., & Hirshman, E. (1998). True and false recognition in MINERVA2: Explanations from a global matching perspective. *Journal of Memory and Language*, *39*, 371–391.

Arndt, J. (2010). The role of memory activation in creating false memories of encoding context. *Journal of Experimental Psychology: Learning Memory and Cognition*, *36*(1), 66–79. https://doi.org/10.1037/a0017394

Arndt, J., & Reder, L. M. (2003). The effect of distinctive visual information on false recognition. *Journal of Memory and Language*, *48*(1), 1–15. https://doi.org/10.1016/S0749-596X(02)00518-1

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*(4), 390–412. https://doi.org/10.1016/j.jml.2007.12.005

Baayen, R. H., Piepenbrock, R., & Gulikers, L. (1995). *The celex Lexical database (Version 2)*. Linguistic Data Consortium, University of Pennsylvania.

Bakker, A., Kirwan, C. B., Miller, M., & Stark, C. E. L. (2008). Pattern separation in the human hippocampal CA3 and dentate gyrus. *Science*, *319*(5870), 1640–1642. https://doi.org/10.1126/science.1152882

Barry, C., Morrison, C. M., & Ellis, A. W. (1997). Naming the Snodgrass and Vanderwart pictures: Effects of age of acquisition, frequency, and name agreement. *Quarterly Journal of Experimental Psychology Section A: Human Experimental Psychology*, *50*(3), 560–585. https://doi.org/10.1080/783663595

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate : A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society*, *57*(1), 289–300.

Bowman, C. R., Chamberlain, J. D., & Dennis, N. A. (2019). Sensory representations supporting memory specificity: Age effects on behavioral and neural discriminability. *Journal of Neuroscience*, *39*(12), 2265–2275. https://doi.org/10.1523/JNEUROSCI.2022-18.2019

Brady, T. F., Konkle, T., Alvarez, G. A., & Oliva, A. (2008). Visual long-term memory has a massive storage capacity for object details. *Proceedings of the National Academy of Sciences*, *105*(38), 14325–14329. https://doi.org/10.1073/pnas.0803390105

Brainerd, C. J., Reyna, V. F., & Mojardin, A. H. (1999). Conjoint recognition. *Psychological Review*, Vol. 106, pp. 160–179. https://doi.org/10.1037//0033-295x.106.1.160

Brainerd, C. J., Reyna, V. F., Wright, R., & Mojardin, A. H. (2003). Recollection rejection : False-memory editing in children and adults. *Psychological Review*, *110*(4), 762–784. https://doi.org/10.1037/0033-295X.110.4.762

Brainerd, C. J., & Wright, R. (2005). Forward association, backward association, and the false-memory illusion. *Journal of Experimental Psychology: Learning Memory and Cognition*, *31*(3), 554–567. https://doi.org/10.1037/0278-7393.31.3.554

Brainerd, C. J., Yang, Y., Reyna, V. F., Howe, M. L., & Mills, B. A. (2008). Semantic

processing in “associative” false memory. *Psychonomic Bulletin & Review*, *15*(6), 1035–1053. https://doi.org/10.3758/PBR.15.6.1035

Brainerd, C. J., & Reyna, V. F. (2002). Fuzzy-trace theory and false memory. *Current Directions in Psychological Science*, *11*(5), 164–169. https://doi.org/10.1111/1467-8721.00192

Brainerd, C. J., & Reyna, V. F. (2005). *The science of false memory*. Oxford University Press.

Brainerd, C. J., Reyna, V. F., & Kneer, R. (1995). False recognition reversal: When similarity is distinctive. *Journal of Memory and Language*, *34*, 157–185.

Brodeur, M. B., Guérard, K., & Bouras, M. (2014). Bank of Standardized Stimuli (BOSS) phase ii: 930 new normative photos. *PLoS ONE*, *9*(9). https://doi.org/10.1371/journal.pone.0106953

Brysbaert, M., & Biemiller, A. (2017). Test-based age-of-acquisition norms for 44 thousand English word meanings. *Behavior Research Methods*, *49*(4), 1520–1523. https://doi.org/10.3758/s13428-016-0811-4

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, *46*(3), 904–911. https://doi.org/10.3758/s13428-013-0403-5

Budson, A. E., Sullivan, A. L., Daffner, K. R., & Schacter, D. L. (2003). Semantic versus phonological false recognition in aging and Alzheimer’s disease. *Brain and Cognition*, *51*(3), 251–261. https://doi.org/10.1016/S0278-2626(03)00030-7

Burnside, K., Hope, C., Gill, E., & Morcom, A. M. (2017). Effects of perceptual similarity but not semantic association on false recognition in aging. *PeerJ*, *5*, e4184. https://doi.org/10.7717/peerj.4184

Cann, D. R., Mcrae, K., & Katz, A. N. (2011). False recall in the Deese-Roediger-McDermott paradigm: The roles of gist and associative strength. *Quarterly Journal of Experimental Psychology*, *64*(8), 1515–1542. https://doi.org/10.1080/17470218.2011.560272

Cann, D. R., McRae, K., & Katz, A. N. (2011). False recall in the Deese–Roediger–McDermott paradigm: The roles of gist and associative strength. *The Quarterly Journal of Experimental Psychology*, *64*(8), 1515–1542. https://doi.org/10.1080/17470218.2011.560272

Chouinard, P. A., & Goodale, M. A. (2012). FMRI-adaptation to highly-rendered color photographs of animals and manipulable artifacts during a classification task. *NeuroImage*, *59*(3), 2941–2951. https://doi.org/10.1016/j.neuroimage.2011.09.073

Clark, H. (1973). The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior*, *12*, 335–359.

Clarke, A., & Tyler, L. K. (2014). Object-specific semantic coding in human perirhinal cortex. *Journal of Neuroscience*, *34*(14), 4766–4775. https://doi.org/10.1523/JNEUROSCI.2828-13.2014

Clarke, A., Devereux, B. J., Randall, B., & Tyler, L. K. (2015). Predicting the time course of individual objects with MEG. *Cerebral Cortex*, *25*(10), 1–11. https://doi.org/10.1093/cercor/bhu203

Coane, J. H., McBride, D. M., Termonen, M. L., & Cutting, J. C. (2016). Categorical and associative relations increase false memory relative to purely associative relations. *Memory and Cognition*, *44*(1), 37–49. https://doi.org/10.3758/s13421-015-0543-1

Coane, J. H., McBride, D. M., & Xu, S. (2020). The feature boost in false memory: The roles of monitoring and critical item identifiability. *Memory*, *28*(4), 481–493. https://doi.org/10.1080/09658211.2020.1735445

De Deyne, S., & Storms, G. (2008). Word associations: Norms for 1,424 Dutch words in a continuous task. *Behavior Research Methods*, *40*(1), 198–205. https://doi.org/10.3758/BRM.40.1.198

DeCarlo, L. T. (1998). Signal detection theory and generalized linear models. *Psychol. Methods*, *3*(2), 186–205. https://doi.org/10.1037/1082-989X.3.2.186

Deese, J. (1959). On the prediction of occurrence of particular verbal intrusions in immediate recall. *Journal of Experimental Psychology*, *58*(1), 17–22. https://doi.org/10.1037/h0046671

Devereux, B. J., Tyler, L. K., Geertzen, J., & Randall, B. (2014). The Centre for Speech, Language and the Brain (CSLB) concept property norms. *Behavior Research Methods*, *46*(4), 1119–1127. https://doi.org/10.3758/s13428-013-0420-4

Forde, E. M. E., & Humphreys, G. W. (1999). Category-specific recognition impairments: A review of important case studies and influential theories. *Aphasiology*, *13*(3), 169–193. https://doi.org/10.1080/026870399402172

Gallo, D. A. (2004). Using recall to reduce false recognition: diagnostic and disqualifying monitoring. *Journal of Experimental Psychology: Learning Memory and Cognition*, *30*(1), 120–128. https://doi.org/10.1037/0278-7393.30.1.120

Gallo, D. A. (2010). False memories and fantastic beliefs: 15 years of the DRM illusion. *Memory & Cognition*, *38*(7), 833–848. https://doi.org/10.3758/MC.38.7.833

Garoff-Eaton, R. J., Slotnick, S. D., & Schacter, D. L. (2006). Not all false memories are created equal: The neural basis of false recognition. *Cerebral Cortex*, *16*(11), 1645–1652. https://doi.org/10.1093/cercor/bhj101

Garrard, P., Lambon Ralph, M. A., Hodges, J. R., & Patterson, K. (2001). Prototypicality , distinctiveness, and intercorrelation: Analyses of the semantic attributes of living and nonliving concepts. *Cognitive Neuropsychology*, *18*(2), 125–174. https://doi.org/10.1080/02643290125857

Glanzer, M., & Adams, J. K. (1985). The mirror effect in recognition memory. *Memory & Cognition*, *13*(1), 8–20. https://doi.org/10.3758/BF03198438

Green, P., & Macleod, C. J. (2016). SIMR: An R package for power analysis of generalized linear mixed models by simulation. *Methods in Ecology and Evolution*, *7*(4), 493–498. https://doi.org/10.1111/2041-210X.12504

Gutchess, A. H., & Schacter, D. L. (2012). The neural correlates of gist-based true and false recognition. *NeuroImage*, *59*(4), 3418–3426. https://doi.org/10.1016/j.neuroimage.2011.11.078

Heit, E., Brockdorff, N., & Lamberts, K. (2003). Adaptive changes of response criterion in recognition memory. *Psychonomic Bulletin and Review*, *10*(3), 718–723. https://doi.org/10.3758/BF03196537

Hintzman, D. L. (1988). Judgments of frequency and recognition memory in a multiple-trace memory model. *Psychological Review*, *95*(4), 528–551. https://doi.org/10.1037/0033-295X.95.4.528

Humphreys, G. W., & Forde, E. M. (2001). Hierarchies, similarity, and interactivity in object

recognition: “category-specific” neuropsychological deficits. *The Behavioral and Brain Sciences*, *24*(3), 453–476; discussion 476-509.

Hunsaker, M. R., & Kesner, R. P. (2013). The operation of pattern separation and pattern completion processes associated with different attributes or domains of memory. *Neuroscience and Biobehavioral Reviews*, *37*(1), 36–58. https://doi.org/10.1016/j.neubiorev.2012.09.014

Kent, C., Lamberts, K., & Patton, R. (2018). Cue quality and criterion setting in recognition memory. *Memory and Cognition*, *46*(5), 757–769. https://doi.org/10.3758/s13421-018-0796-6

Kleiner, M., Brainard, D., Pelli, D., Ingling, A., Murray, R., & Broussard, C. (2007). What's new in psychtoolbox-3. *Perception*, *36*(14), 1-16.

Konkle, T., Brady, T. F., Alvarez, G. A., & Oliva, A. (2010). Conceptual distinctiveness supports detailed visual long-term memory for real-world objects. *Journal of Experimental Psychology: General*, *139*(3), 558–578. https://doi.org/10.1037/a0019165

Koustaal, W., Schacter, D. L., Galluccio, L., & Stofer, K. A. (1999). Reducing gist-based false recognition in older adults: Encoding and retrieval manipulations. *Psychology and Aging*, *14*(2), 220–237. https://doi.org/10.1037/0882-7974.14.2.220

Koutstaal, W., Reddy, C., Jackson, E. M., Prince, S., Cendan, D. L., & Schacter, D. L. (2003). False recognition of abstract versus common objects in older and younger adults: Testing the semantic categorization account. *Journal of Experimental Psychology: Learning Memory and Cognition*, *29*(4), 499–510. https://doi.org/10.1037/0278-7393.29.4.499

Koutstaal, W., & Schacter, D. L. (1997). Gist-based false recognition of pictures in older and younger adults. *Journal of Memory and Language*, *583*(37), 555–583.

Lacy, J. W., Yassa, M. A., Stark, S. M., Muftuler, L. T., & Stark, C. E. L. (2011). Distinct pattern separation related transfer functions in human CA3/dentate and CA1 revealed using highresolution fMRI and variable mnemonic similarity. *Learning and Memory*, *18*(1), 15–18. https://doi.org/10.1101/lm.1971111

Loiotile, R. E., & Courtney, S. M. (2015). A signal detection theory analysis of behavioral pattern separation paradigms. *Learning and Memory*, *22*(8), 364–369. https://doi.org/10.1101/lm.038141.115

Marr, D. (1971). Simple memory: A theory for archicortex. *Philosophical Transactions of the Royal Society of London*, *262*, 23- 81.

Matuschek, H., Kliegl, R., Vasishth, S., Baayen, H., & Bates, D. (2017). Balancing Type I error and power in linear mixed models. *Journal of Memory and Language*, *94*, 305–315. https://doi.org/10.1016/j.jml.2017.01.001

McClelland, J. L., McNaughton, B. L., & O’Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review*, *102*(3), 419–457. https://doi.org/10.1037/0033-295X.102.3.419

McNaughton, B. L., & Morris, R. G. M. (1987). Hippocampal synaptic enhancement and information storage. *Trends in Neural Sciences*, *10*(10), 408–415.

McRae, K., Cree, G. S., Seidenberg, M. S., & McNorgan, C. (2005). Semantic feature production norms for a large set of living and nonliving things. *Behavior Research Methods*, *37*(4), 547–559. https://doi.org/10.3758/BF03192726

McRae, K., De Sa, V. R., & Seidenberg, M. S. (1997). On the nature and scope of featural representations of word meaning. *Journal of Experimental Psychology: General*, *126*(2), 99–130. https://doi.org/10.1037/0096-3445.126.2.99

Migo, E., Montaldi, D., Norman, K. A., Quamme, J., & Mayes, A. (2009). The contribution of familiarity to recognition memory is a function of test format when using similar foils. *Quarterly Journal of Experimental Psychology*, *62*(6), 1198–1215. https://doi.org/10.1080/17470210802391599

Montefinese, M., Ambrosini, E., Fairfield, B., & Mammarella, N. (2013). Semantic memory : A feature-based analysis and new norms for Italian. *Behavior Research Methods*, *45*(2), 440–461. https://doi.org/10.3758/s13428-012-0263-4

Montefinese, M., Vinson, D., & Ambrosini, E. (2018). Recognition memory and featural similarity between concepts: The pupil's point of view. *Biological Psychology*, *135*, 159–169. https://doi.org/10.1016/j.biopsycho.2018.04.004

Montefinese, M., Zannino, G. D., & Ambrosini, E. (2015). Semantic similarity between old and new items produces false alarms in recognition memory. *Psychological Research*, *79*(5), 785–794. https://doi.org/10.1007/s00426-014-0615-z

Moss, H. E., Tyler, L. K., & Jennings, F. (1997). When leopards lose their spots: Knowledge of visual properties in category-specific deficits for living things. *Cognitive Neuropsychology*, *14*(6), 901–950. https://doi.org/10.1080/026432997381394

Motley, S. E., & Kirwan, C. B. (2012). A parametric investigation of pattern separation processes in the medial temporal lobe. *Journal of Neuroscience*, *32*(38), 13076–13084. https://doi.org/10.1523/JNEUROSCI.5920-11.2012

Nelson, D., McEvoy, C., & Schreiber, T. (2004). The University of South Florida word association, rhyme, and word fragment norms. *Behavior Research Methods, Instruments, & Computers*, *36*(3), 402–407

Norman, K. A., & O'Reilly, R. C. (2003). Modeling hippocampal and noecortical contributions to recognition memory: A complementary-learning systems approach. *Psychological Review*, *110*(4), 611–646

Norman, K. A. (2010). How hippocampus and cortex contribute to recognition memory: Revisiting the complementary learning systems model. *Hippocampus*, *20*(11), 1217–1227. https://doi.org/10.1002/hipo.20855

Oldfield, R. C., & Wingfield, A. (1964). The time it takes to name an object. *Nature*, *202*, 1031–1032

Pidgeon, L. M., & Morcom, A. M. (2014). Age-related increases in false recognition: The role of perceptual and conceptual similarity. *Frontiers in Aging Neuroscience*, *6*, 1–17. https://doi.org/10.3389/fnagi.2014.00283

Pidgeon, L. M., & Morcom, A. M. (2016). Cortical pattern separation and item-specific memory encoding. *Neuropsychologia*, *85*, 256–271. https://doi.org/10.1016/j.neuropsychologia.2016.03.026

Quené, H., & van den Bergh, H. (2008). Examples of mixed-effects modeling with crossed random effects and with binomial data. *Journal of Memory and Language*, *59*(4), 413–425. https://doi.org/10.1016/j.jml.2008.02.002

Randall, B., Moss, H. E., Rodd, J. M., Greer, M., & Tyler, L. K. (2004). Distinctiveness and correlation in conceptual structure: Behavioral and computational studies. *Journal of*

Experimental Psychology: Learning Memory and Cognition, *30*(2), 393–406. https://doi.org/10.1037/0278-7393.30.2.393

Reagh, Z. M., Ho, H. D., Leal, S. L., Noche, J. A., Chun, A., Murray, E. A., & Yassa, M. A. (2016). Greater loss of object than spatial mnemonic discrimination in aged adults. *Hippocampus*, *26*, 417–422. https://doi.org/10.1002/hipo.22562

Reagh, Z. M., Roberts, J. M., Ly, M., Diprospero, N., Murray, E., & Yassa, M. A. (2014). Spatial discrimination deficits as a function of mnemonic interference in aged adults with and without memory impairment. *Hippocampus*, *24*(3), 303–314. https://doi.org/10.1002/hipo.22224

Reppa, I., Williams, K. E., Greville, W. J., & Saunders, J. (2020). The relative contribution of shape and colour to object memory. *Memory and Cognition*, *48*(8), 1504–1521. https://doi.org/10.3758/s13421-020-01058-w

Reyna, V. F., & Brainerd, C. J. (1995). Fuzzy-trace theory: An interim synthesis. *Learning and Individual Differences*, *7*(1), 1–75. https://doi.org/10.1016/1041-6080(95)90031-4

Reyna, V. F., Corbin, J. C., Weldon, R. B., & Brainerd, C. J. (2016). How fuzzy-trace theory predicts true and false memories for words, sentences, and narratives. *Journal of Applied Research in Memory and Cognition*, *5*(1), 1–9. https://doi.org/10.1016/j.jarmac.2015.12.003

Reyna, V. F., & Kiernan, B. (1994). Development of gist versus verbatim memory in sentence recognition: Effects of lexical familiarity, semantic content, encoding instructions, and retention interval. *Developmental Psychology*, *30*(2), 178–191

Riesenhuber, M., & Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nature Neuroscience*, *2*(11), 1019–1025

Roediger, H. L., & McDermott, K. B. (1995). Creating false memories: Remembering words not presented in lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*(4), 803–814

Roediger, H. L., & McDermott, K. B. (2000). Tricks of memory. *Current Directions in Psychological Science*, *9*(4), 123–127. https://doi.org/10.1111/1467-8721.00075

Roediger, H. L., Watson, J. M., McDermott, K. B., & Gallo, D. A. (2001). Factors that determine false recall: A multiple regression analysis. *Psychonomic Bulletin & Review*, *8*(3), 385–407. https://doi.org/10.3758/BF03196177

Rosch, E., & Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology*, *7*(4), 573–605. https://doi.org/10.1016/0010-0285(75)90024-9

Rubner, Y., Tomasi, C., & Guibas, L. J. (2000). The earth mover's distance as a metric for image retrieval. *International Journal of Computer Vision*, *40*(2), 99–121.

Schmidt, S. R. (1996). Category typicality effects in episodic memory: Testing models of distinctiveness. *Memory and Cognition*, *24*(5), 595–607. https://doi.org/10.3758/BF03201086

Seamon, J. G., Luo, C. R., Schlegel, S. E., Greene, S. E., & Goldenberg, A. B. (2000). False memory for categorized pictures and words: the category associates procedure for studying memory errors in children and adults. *Journal of Memory and Language*, *42*, 120–146.

Serre, T., Wolf, L., Bileschi, S., Riesenhuber, M., & Poggio, T. (2007). Robust object

recognition with cortex-like mechanisms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *29*(3), 411–426. https://doi.org/10.1109/TPAMI.2007.56

Slotnick, S. D., & Schacter, D. L. (2004). A sensory signature that distinguishes true from false memories. *Nature Neuroscience*, *7*(6), 664–672. https://doi.org/10.1038/nn1252

Snodgrass, J. G., & Vanderwart, M. (1980). A standardized set of 260 pictures: Norms for name agreement, image agreement, familiarity, and visual complexity. *Journal of Experimental Psychology: Human Learning & Memory*, *6*(2), 174–215. https://doi.org/10.1037/0278-7393.6.2.174

Stark, S. M., Yassa, M. A., Lacy, J. W., & Stark, C. E. L. (2013). A task to assess behavioral pattern separation (BPS) in humans: Data from healthy aging and mild cognitive impairment. *Neuropsychologia*, *51*(12), 2442–2449. https://doi.org/10.1016/j.neuropsychologia.2012.12.014

Taylor, K. I., Devereux, B. J., Acres, K., Randall, B., & Tyler, L. K. (2012). Contrasting effects of feature-based statistics on the categorisation and basic-level identification of visual objects. *Cognition*, *122*(3), 363–374. https://doi.org/10.1016/j.cognition.2011.11.001

Taylor, K. I., Devereux, B. J., & Tyler, L. K. (2011). Conceptual structure: Towards an integrated neurocognitive account. *Language and Cognitive Processes*, *26*(9), 1368–1401. https://doi.org/10.1080/01690965.2011.568227

Tyler, L. K., & Moss, H. E. (2001). Towards a distributed account of conceptual knowledge. *Trends in Cognitive Sciences*, *5*(6), 244–252. https://doi.org/10.1016/S1364-6613(00)01651-X

Tyler, L. K., Stamatakis, E. A., Bright, P., Acres, K., Abdallah, S., Rodd, J. M., & Moss, H. E. (2004). Processing objects at different levels of specificity. *Journal of Cognitive Neuroscience*, *16*(3), 351–362. https://doi.org/10.1162/089892904322926692

van Heuven, W. J. B., Mandera, P., Keuleers, E., & Brysbaert, M. (2014). SUBTLEX-UK: A new and improved word frequency database for British English. *Quarterly Journal of Experimental Psychology*, *67*(6), 1176–1190. https://doi.org/10.1080/17470218.2013.850521

Vigliocco, G., Vinson, D. P., Lewis, W., & Garrett, M. F. (2004). Representing the meanings of object and action words: The featural and unitary semantic space hypothesis. *Cognitive Psychology*, *48*(4), 422–488. https://doi.org/10.1016/j.cogpsych.2003.09.001

Voss, J. L., Lucas, H. D., & Paller, K. A. (2012). More than a feeling: pervasive influences of memory without awareness of retrieval. *Cognitive Neuroscience*, *3*(3–4), 193–207. https://doi.org/10.1080/17588928.2012.674935

Warrington, E. K., & Shallice, T. (1984). Category specific semantic impairments. *Brain*, *107*(3), 829–853. https://doi.org/10.1093/brain/107.3.829

Watson, J. M., Balota, D. A., & Roediger, H. L. (2003). Creating false memories with hybrid lists of semantic and phonological associates: Over-additive false memories produced by converging associative networks. *Journal of Memory and Language*, *49*(1), 95–118. https://doi.org/10.1016/S0749-596X(03)00019-6

Watson, J. M., Balota, D. A., & Sergent-Marshall, S. D. (2001). Semantic, phonological, and hybrid veridical and false memories in healthy older adults and in individuals with dementia of the Alzheimer type. *Neuropsychology*, *15*(2), 254–267. https://doi.org/10.1037/0894-4105.15.2.254

Weller, H. I., & Westneat, M. W. (2019). Quantitative color profiling of digital images with earth mover's distance using the R package colordistance. *PeerJ*, *7*, 1–31. https://doi.org/10.7717/peerj.6398

Wilson, I. A., Gallagher, M., Eichenbaum, H., & Tanila, H. (2006). Neurocognitive aging: prior memories hinder new hippocampal encoding. *Trends in Neurosciences*, *29*(12), 662–670. https://doi.org/10.1016/j.tins.2006.10.002

Wright, D. B., Horry, R., & Skagerberg, E. M. (2009). Functions for traditional and multilevel approaches to signal detection theory. *Behavior Research Methods*, *41*(2), 257–267. https://doi.org/10.3758/BRM.41.2.257

Yassa, M. A., Lacy, J. W., Stark, S. M., Albert, M. S., Gallagher, M., & Stark, C. E. L. (2011). Pattern separation deficits associated with increased hippocampal CA3 and dentate gyrus activity in nondemented older adults. *Hippocampus*, *21*(9), 968–979. https://doi.org/10.1002/hipo.20808

Yassa, M. A., & Stark, C. E. L. (2011). Pattern separation in the hippocampus. *Trends in Neurosciences*, *34*(10), 515–525. https://doi.org/10.1016/j.tins.2011.06.006

Yotsumoto, Y., Kahana, M. J., Wilson, H. R., & Sekuler, R. (2007). Recognition memory for realistic synthetic faces. *Memory and Cognition*, *35*(6), 1233–1244. https://doi.org/10.3758/BF03193597

Supplemental Materials

Nuisance Variables

Word Associations

The strength of associations between words is an established modifier of mnemonic discrimination (Deese, 1959; Roediger et al., 2001). Both forward associative strength (FAS, from an unstudied lure concept to a studied concept) and backward associative strength (BAS, from a studied concept to an unstudied lure concept) may contribute to misrecognition and false recall of lures (Brainerd & Wright, 2005; Roediger et al., 2001). However, the current English language norms (Nelson et al., 1998) only provide data for 156 of the 200 concepts with semantic feature norms used in this study (Devereux et al., 2014). The Nelson et al. (1998) norms were also gathered decades ago (from 1973) in the US. To ensure that association scores were relevant for our participants, we gathered our own association data using a method similar to De Deyne and Storms (2008).

Participants

Two hundred and six participants contributed (Age: M = 24.7, SD = 8.7, 147 female, 59 male). Participants were recruited using social media and Mechanical Turk. They were required to live in the UK and be aged 18-50 years. Non-native English speakers self-evaluated their English ability using a scale from 0 (none) to 5 (fluent) on four dimensions: expression, comprehension, reading, and writing. A score of at least 16 out of 20 was required. Forty further participants were removed who did not meet these criteria. Each concept was seen by an average of 32.4 participants (SD = 2.5).

Stimuli and Procedure

Participants each rated 30 randomly chosen concepts from the 200 used in the study. For each concept, participants were asked to list the first three words that came to mind when reading the name of the concept, ranked such that the most prevalent (or first thought of) was listed first. Participants could also mark a word as unknown or provide less than three answers. There was no time limit to complete the task.

Calculation of Mean BAS and Mean FAS

Associations for a concept were manually sorted and counted to take account of different spellings, spelling mistakes, and capitalizations. Following Nelson et al. (2004), conjugated words and plural/singular words were transformed into the most frequent instance. Following De Deyne and Storms (2008), idiosyncratic words (associations only produced once) were removed in the calculation of frequencies. Then, the associations were tallied for each concept and divided by the number of associations produced for that concept to create relative frequencies of each association.

A useful metric is how much on average a specific concept is associated with the other concepts used in the study (De Deyne & Storms, 2008). This is referred to as mean backward associative

strength (MBAS) and mean forward associative strength (MFAS). To calculate these, we followed procedures used by De Deyne and Storms (2008), and Montefinese, Zannino, and Ambrosini (2015). Association data were entered in a 200 x 200 matrix in which rows corresponded to concepts and columns to their generated associations with other concepts. We entered each association frequency in the corresponding cell. The MFAS was then calculated by averaging over frequencies in a given concept's row, and MBAS by averaging over a given concept's column.

Word frequency

Frequently encountered words are processed faster than words that are rarely encountered (Oldfield & Wingfield, 1964). Low-frequency words are correctly recognised and correctly rejected more often than high-frequency words (Glanzer & Adams, 1985). We used log-transformed word frequencies from SUBTTLEX-UK, based on British English television subtitles (van Heuven et al., 2014).

Concreteness

Concreteness refers to the extent to which the concept denoted by a word refers to a perceptible entity (Paivio & Begg, 1971). More concrete words are easier to remember than more abstract words (Gorman, 1961; Paivio, 2013). We used norms from Brysbaert et al. (2014). Participants rated how concrete the meaning of each word was by using a 5-point scale.

Age of Acquisition

Age of acquisition (AoA) is known to influence memory: words learned early in life are less well recognized (Dewhurst et al., 1998). We used Brysbaert & Biemille's (2017) update of Dale and O'Rourke's (1981) norms.

Phonological Neighborhood Density

The phonological neighbourhood density (PND) of a word is the number of words in the language which differ only by the addition, subtraction or substitution of a single phoneme from the target word (Luce & Pisoni, 1998). Items with high PND tend to be processed less quickly and/or accurately. We used norms from the CELEX database (Baayen et al., 1995).

Visual Complexity

The visual complexity measures reflect superficial visual characteristics of images. More complex stimuli may be more easily recognized (Snodgrass & Vanderwart, 1980). We included two different measures of image complexity: the number of non-white pixels in the image, and color entropy, a measure of the color variability of an image. Images with a large proportion of pixels sharing the same color should be less visually complex (Chouinard & Goodale, 2012). Color entropy was computed by finding the relative frequency of all colors that occur in the non-white pixels in the image and calculating the entropy of this probability distribution.

Concept Familiarity

We included concept familiarity as an attribute of images. In recognition memory, more unfamiliar pictures are better recognised than familiar pictures (Snodgrass & Vanderwart, 1980).

Each covariate we considered in the study was standardized and the statistics that follow were calculated over the new set of z-score values.

Supplemental Results

Experiment 1

Figure 4

Schematic Depiction of Additional Graded Perceptual Confusability Measures

Note. Rows show individual exemplars with the highest (right side) and lowest (left side) scores on each metric of perceptual confusability. The graded perceptual confusability measures define confusability indexing an image's overall similarity to the full set of images. C1 and C2 were obtained from gray-scaled version of the images depicted in Figure 4. For definitions see Variables of Interest section.

Table 5

Principal Component Analysis of Nuisance Variables in Experiment 1

	Components						
	PC1	PC2	PC3	PC4	PC5	PC6	PC7
Nb of non-white pixel					0.997		
Color entropy	0.154		0.701			-0.272	0.264
Mean BAS	-0.536						-0.334
Mean FAS	-0.811						0.272
PND							-0.822
Concept familiarity	-0.156		0.710			0.258	-0.243
Word co-occurrence				0.992			
Concreteness		-0.700					
Age of acquisition						0.921	
Word Frequency		-0.708					

Note. The top 7 principal components from the PCA with varimax rotation of nuisance variables. Cumulative variance explained is 86.14 %.

Table 6

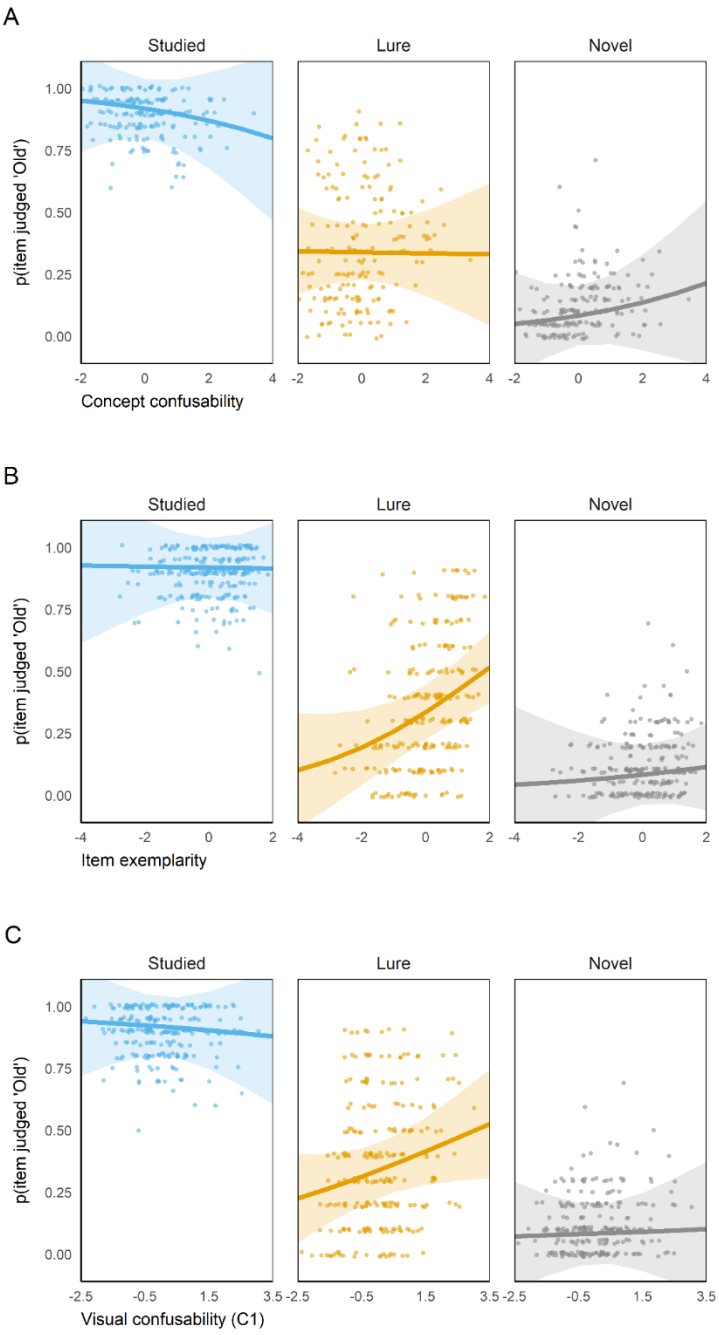
Model Selection for Experiment 1

Model	AICc	ΔAICc	AICcWt	Cum.Wt	LL
Concept + Item + PCs	8821.92	0.00	0.98	0.98	-4380.87
Concept + Item	8830.05	8.13	0.02	1.00	-4391.97
Item + PCs	8876.02	54.10	0.00	1.00	-4413.95
Item	8886.94	65.02	0.00	1.00	-4426.44
Concept + PCs	8899.91	77.98	0.00	1.00	-4431.92
Concept	8911.89	89.96	0.00	1.00	-4444.93
PCs	14420.23	5598.30	0.00	1.00	-7200.10
Null	14430.75	5608.83	0.00	1.00	-7212.38

Note. Summary of AICc results for models including concept-level, item-level, and confounds principal component variables; PCs = Principal components.

Figure 5

Effects of Semantic and Perceptual Variables on Raw Recognition Measures in Experiment 1



Note. Plots show effects of semantic and perceptual variables on raw recognition responses by item type. The plot lines represent the raw probabilities of endorsing studied items as "old" (light blue), lures as "old" (orange), and novel items as "old" (grey). Panel A, B, and C show the effects of concept confusability, item exemplarity, and C1 visual confusability in Experiment 1. Note that for concept-level variables (concept confusability) there are data points for each concept, and for item-level variables (item exemplarity and C1 visual confusability) there are data points for each exemplar image. The clustering around discrete values of p("old") reflects the small numbers of observations for individual exemplars (see Experiment 1, Materials and Methods, Stimuli).

Table 7

Results of Experiment 1 Using Graded Perceptual Confusability Metrics

Variable	Estimate	*d'*	SE	z-value	*p*
(Intercept)	-2.42	-1.38	0.12	-20.09	<.001
Lure	1.73	0.96	0.07	23.37	<.001
Studied	4.86	2.75	0.09	52.24	<.001
Number of Features	0.15	0.08	0.08	1.94	.084
Concept Confusability	0.29	0.16	0.08	3.61	<.001
Visual Confusability (C1)	-0.14	-0.08	0.07	-1.89	.084
Visual Confusability (C2)	0.14	0.06	0.08	1.67	.117
Color Confusability	0.13	0.07	0.08	1.69	.117
Item Exemplarity	0.18	0.10	0.07	2.61	.022
Lure × Number of Features	-0.32	-0.18	0.07	-4.43	<.001
Studied × Number of Features	-0.12	-0.06	0.09	-1.40	.190
Lure × Concept Confusability	-0.27	-0.14	0.07	-3.74	<.001
Studied × Concept Confusability	-0.57	-0.30	0.08	-6.73	<.001
Lure × Visual Confusability (C1)	0.14	0.08	0.08	1.88	.084
Studied × Visual Confusability (C1)	0.20	0.11	0.09	2.27	.041
Lure × Visual Confusability (C2)	-0.08	-0.03	0.08	-0.92	.355
Studied × Visual Confusability (C2)	-0.11	-0.05	0.10	-1.11	.297
Lure × Color Confusability	-0.08	-0.04	0.07	-1.05	.307
Studied × Color Confusability	-0.26	-0.14	0.09	-3.00	.007
Lure × Item Exemplarity	0.19	0.12	0.08	2.44	.028
Studied × Item Exemplarity	-0.23	-0.11	0.09	-2.52	.025

Note. The reference level of condition is set to "novel". Parameter estimates (logOR), *d'* equivalent, standard errors, z-values, and FDR-corrected *p*-values are listed for condition, concept-level, and item-level variables in the winning (full) linear mixed model selected with AIC. Graded perceptual confusability measures were reported in the model above. See Material and Methods, and Variables of Interest for details. SE = Standard Error.

Experiment 2

Table 8

Principal Component Analysis of Nuisance Variables in Experiment 2

	Components						
	PC1	PC2	PC3	PC4	PC5	PC6	PC7
Nb of non-white pixel	-0.203		-0.675	-0.116		-0.229	0.112
Color entropy					-0.989		
Mean BAS	0.602						-0.238
Mean FAS	0.744						0.257
PND							-0.924
Familiarity	0.198		-0.733	0.108		-0.207	
Word co-occurrence				0.992			
Concreteness		0.706					
Age of acquisition						-0.947	
Word frequency		0.707					

Note. The top 7 principal components from the PCA with varimax rotation of nuisance variables. Cumulative variance explained is 86.54 %.

Table 9

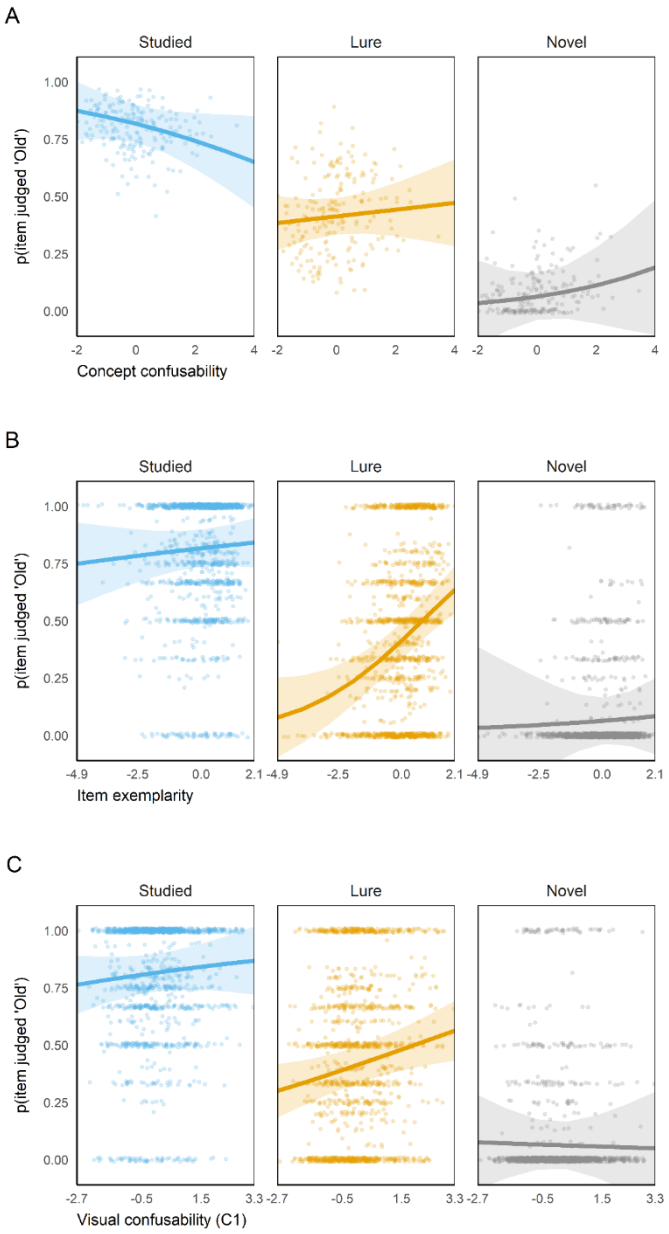
Model Selection for Experiment 2

Model	AICc	ΔAICc	AICcWt	Cum.Wt	LL
Concept + Item + PCs	17597.96	0.00	0.89	0.89	-8768.93
Concept + Item	17602.07	4.12	0.11	1.00	-8778.00
Item + PCs	17660.69	62.73	0.00	1.00	-8806.31
Item	17665.99	68.03	0.00	1.00	-8815.98
Concept + PCs	17854.98	257.02	0.00	1.00	-8909.47
Concept	17862.95	264.99	0.00	1.00	-8920.47
PCs	23836.19	6238.23	0.00	1.00	-11908.09
Null	23856.48	6258.53	0.00	1.00	-11925.24

Note. Summary of AICc results for models including concept-level, item-level, and confounds principal component variables; PCs = Principal components.

Figure 6

Effects of Semantic and Perceptual Variables on Raw Recognition Measures in Experiment 2



Note. Plots show effects of semantic and perceptual variables on raw recognition responses by item type. The plot lines represent the raw probabilities of endorsing studied items as "old" (light blue), lures as "old" (orange), and novel items as "old" (grey). Panel A, B, and C show the effects of concept confusability, item exemplarity, and C1 visual confusability in Experiment 2. Note that for concept-level variables (concept confusability) there are data points for each concept, and for item-level variables (item exemplarity and C1 visual confusability) there are data points for each exemplar image. The clustering around discrete values of p("old") reflects the small numbers of observations for individual exemplars, as 300 of 1800 images were randomly allocated to the 3 test conditions for each participant (see Experiment 2, Materials and Methods, Stimuli).

Table 10

Results Including Study Set Size Effects for Experiment 2

Variable	Estimate	*d'*	SE	z-value	*p*
(Intercept)	-0.86	-0.51	0.09	-9.94	<.001
Novel	-1.85	-1.03	0.08	-23.48	<.001
Lure-Set8	0.98	0.59	0.05	17.90	<.001
Studied-Set2	2.18	1.30	0.06	37.26	<.001
Studied-Set8	2.57	1.53	0.06	40.54	<.001
Number of Features	-0.02	-0.01	0.05	-0.30	.890
Concept Confusability	0.09	0.06	0.05	1.71	.162
Visual Confusability (C1)	0.13	0.07	0.04	2.94	.010
Visual Confusability (C2)	0.06	0.03	0.05	1.31	.304
Color Confusability	0.08	0.05	0.04	1.81	.135
Item Exemplarity	0.43	0.25	0.05	9.52	<.001
Novel × Number of Features	0.01	0.00	0.08	0.11	.941
Lure-Set8 × Number of Features	-0.01	0.00	0.06	-0.11	.941
Studied-Set2 × Number of Features	-0.03	-0.02	0.06	-0.45	.813
Studied-Set8 × Number of Features	-0.16	-0.09	0.07	-2.28	.050
Novel × Concept Confusability	0.21	0.10	0.08	2.81	.013
Lure-Set8 × Concept Confusability	-0.08	-0.05	0.06	-1.37	.286
Studied-Set2 × Concept Confusability	-0.31	-0.19	0.06	-5.28	<.001
Studied-Set8 × Concept Confusability	-0.33	-0.19	0.06	-5.12	<.001
Novel × Visual Confusability (C1)	-0.20	-0.10	0.08	-2.54	.026
Lure-Set8 × Visual Confusability (C1)	0.14	0.09	0.06	2.53	.026
Studied-Set2 × Visual Confusability (C1)	0.02	0.01	0.06	0.27	.890
Studied-Set8 × Visual Confusability (C1)	-0.03	-0.01	0.06	-0.42	.813
Novel × Visual Confusability (C2)	0.16	0.07	0.09	1.88	.123
Lure-Set8 × Visual Confusability (C2)	0.01	0.01	0.06	0.16	.941
Studied-Set2 × Visual Confusability (C2)	-0.05	-0.03	0.06	-0.92	.503
Studied-Set8 × Visual Confusability (C2)	0.05	0.03	0.06	0.84	.539
Novel × Color Confusability	0.00	-0.01	0.08	-0.01	.988
Lure-Set8 × Color Confusability	0.06	0.03	0.06	1.03	.440
Studied-Set2 × Color Confusability	-0.06	-0.04	0.06	-1.09	.255
Studied-Set8 × Color Confusability	-0.09	-0.05	0.06	-1.46	.417
Novel × Item Exemplarity	-0.28	-0.17	0.08	-3.54	.001
Lure-Set8 × Item Exemplarity	0.04	0.03	0.06	0.63	.689
Studied-Set2 × Item Exemplarity	-0.38	-0.22	0.06	-6.22	<.001
Studied-Set8 × Item Exemplarity	-0.30	-0.18	0.06	-4.83	<.001

Note. The reference level of condition is set to "Lure-Set2". Parameter estimates (logOR), *d'* equivalent, standard errors, z-values, and FDR-corrected *p*-values are listed for condition, concept-level, and item-level variables in the winning (full) linear mixed model selected with AIC. Nearest neighbour perceptual confusability measures were reported in the model above. See Material and Methods, and Variables of Interest and for details. SE = Standard Error.

Table 11

Results of Experiment 2 Using Graded Perceptual Confusability Metrics

Variable	Estimate	*d'*	SE	z-value	*p*
(Intercept)	-2.71	-1.55	0.10	-26.66	<.001
Lure	2.36	1.34	0.07	31.98	<.001
Studied	4.20	2.44	0.08	54.11	<.001
Number of Features	-0.01	-0.01	0.07	-0.16	.917
Concept Confusability	0.29	0.15	0.07	4.05	<.001
Visual Confusability (C1)	0.06	0.03	0.07	0.80	.589
Visual Confusability (C2)	0.30	0.14	0.08	3.96	<.001
Color Confusability	0.02	0.00	0.07	0.30	.896
Item Exemplarity	0.17	0.09	0.07	2.41	.030
Lure × Number of Features	0.00	0.00	0.07	-0.04	.969
Studied × Number of Features	-0.08	-0.05	0.07	-1.10	.405
Lure × Concept Confusability	-0.22	-0.10	0.07	-3.09	.005
Studied × Concept Confusability	-0.51	-0.27	0.07	-6.97	<.001
Lure × Visual Confusability (C1)	-0.05	-0.03	0.07	-0.75	.597
Studied × Visual Confusability (C1)	-0.02	0.00	0.08	-0.23	.908
Lure × Visual Confusability (C2)	-0.25	-0.11	0.08	-3.12	.005
Studied × Visual Confusability (C2)	-0.23	-0.10	0.08	-2.84	.009
Lure × Color Confusability	-0.04	-0.01	0.07	-0.57	.699
Studied × Color Confusability	-0.11	-0.05	0.07	-1.54	.218
Lure × Item Exemplarity	0.27	0.18	0.07	3.68	<.001
Studied × Item Exemplarity	-0.09	-0.04	0.08	-1.17	.390

Note. The reference level of condition is set to "novel". Parameter estimates (logOR), *d'* equivalent, standard errors, z-values, and FDR-corrected *p*-values are listed for condition, concept-level, and item-level variables in the winning (full) linear mixed model selected with AIC. Graded perceptual confusability measures were reported in the model above. See Material and Methods, and Variables of Interest for details. SE = Standard Error.

Experiment 1 and Experiment 2
Table 12

Results for Between-Experiment comparison

Variable	Estimate	*d'*	SE	z-value	*p*
(Intercept)	-2.31	-1.33	0.10	-22.25	<.001
Experiment 2	-0.39	-0.21	0.14	-2.71	.023
Experiment 2 × Lure	0.71	0.40	0.10	6.93	<.001
Experiment 2 × Studied	-0.45	-0.24	0.11	-3.96	<.001
Experiment 2 × Number of Features	-0.12	-0.07	0.09	-1.35	.322
Experiment 2 × Concept Confusability	0.05	0.02	0.09	0.59	.725
Experiment 2 × Visual Confusability (C1)	-0.15	-0.07	0.09	-1.74	.174
Experiment 2 × Visual Confusability (C2)	0.08	0.03	0.11	0.79	.604
Experiment 2 × Color Confusability	0.05	0.01	0.10	0.50	.766
Experiment 2 × Item Exemplarity	-0.06	-0.02	0.09	-0.65	.700
Experiment 2 × Lure × Number of Features	0.29	0.17	0.10	2.97	.011
Experiment 2 × Studied × Number of Features	0.02	0.00	0.11	0.14	.907
Experiment 2 × Lure × Concept Confusability	0.02	0.02	0.10	0.25	.872
Experiment 2 × Studied × Concept Confusability	-0.02	-0.01	0.11	-0.21	.872
Experiment 2 × Lure × Visual Confusability (C1)	0.09	0.04	0.10	0.91	.522
Experiment 2 × Studied × Visual Confusability (C1)	0.37	0.20	0.11	3.42	.003
Experiment 2 × Lure × Visual Confusability (C2)	-0.03	0.00	0.12	-0.28	.872
Experiment 2 × Studied × Visual Confusability (C2)	-0.06	-0.01	0.13	-0.47	.766
Experiment 2 × Lure × Color Confusability	0.05	0.04	0.11	0.49	.766
Experiment 2 × Studied × Color Confusability	0.03	0.03	0.12	0.24	.872
Experiment 2 × Lure × Item Exemplarity	0.11	0.06	0.10	1.05	.444
Experiment 2 × Studied × Item Exemplarity	0.13	0.07	0.11	1.19	.408

Note. The reference levels of condition and experiment are set to "novel" and "Experiment 1". Parameter estimates (logOR), *d'* equivalent, standard errors, z-values, and FDR-corrected *p*-values are listed for condition, concept-level, and item-level variables in the winning (full) linear mixed model selected with AIC. Nearest neighbor perceptual confusability measures were used in the model above. See Material and Methods, and Variables of Interest for details. SE = Standard Error.

References

Baayen, R. H., Piepenbrock, R., & Gulikers, L. (1995). *The celex Lexical database (Version 2)*. Linguistic Data Consortium, University of Pennsylvania.

Brainerd, C. J., & Wright, R. (2005). Forward association, backward association, and the false-memory Illusion. *Journal of Experimental Psychology: Learning Memory and Cognition*, *31*(3), 554–567. https://doi.org/10.1037/0278-7393.31.3.554.

Brysbaert, M., & Biemiller, A. (2017). Test-based age-of-acquisition norms for 44 thousand English word meanings. *Behavior Research Methods*, *49*(4), 1520–1523. https://doi.org/10.3758/s13428-016-0811-4.

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, *46*(3), 904–911. https://doi.org/10.3758/s13428-013-0403-5.

Chouinard, P. A., & Goodale, M. A. (2012). FMRI-adaptation to highly-rendered color photographs of animals and manipulable artifacts during a classification task. *NeuroImage*, *59*(3), 2941–2951. https://doi.org/10.1016/j.neuroimage.2011.09.073.

Dale, E., & O'Rourke, J. (1981). *The living word vocabulary*. World Book Childcraft International.

De Deyne, S., & Storms, G. (2008). Word associations: Norms for 1,424 Dutch words in a continuous task. *Behavior Research Methods*, *40*(1), 198–205. https://doi.org/10.3758/BRM.40.1.198.

Deese, J. (1959). On the prediction of occurrence of particular verbal intrusions in immediate recall. *Journal of Experimental Psychology*, *58*(1), 17–22. https://doi.org/10.1037/h0046671.

Devereux, B. J., Tyler, L. K., Geertzen, J., & Randall, B. (2014). The Centre for Speech, Language and the Brain (CSLB) concept property norms. *Behavior Research Methods*, *46*(4), 1119–1127. https://doi.org/10.3758/s13428-013-0420-4.

Dewhurst, S. A., Hitch, G., & Barry, C. (1998). Separate effects of word frequency and age of acquisition in recogntion and recall. *Journal of Experimental Psychology*, *24*(2), 284–298.

Glanzer, M., & Adams, J. K. (1985). The mirror effect in recognition memory. *Memory & Cognition*, *13*(1), 8–20. https://doi.org/10.3758/BF03198438.

Gorman, A. M. (1961). Recognition memory for nouns as a function of abstractness and frequency. *Journal of Experimental Psychology*, *61*(1), 23–29. https://doi.org/10.1037/h0040561.

Luce, P. A., & Pisoni, D. B. (1998). Recognizing spoken words: The neighborhood activation model. *Ear Hearing*, *19*(1), 1–36.

Montefinese, M., Zannino, G. D., & Ambrosini, E. (2015). Semantic similarity between old and new items produces false alarms in recognition memory. *Psychological Research*, *79*(5), 785–794. https://doi.org/10.1007/s00426-014-0615-z.

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (1998). The University of South Florida word association, rhyme, and word fragment norms. http://www.usf.edu/FreeAssociation/.

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (2004). The University of South Florida word association, rhyme, and word fragment norms. *Behavior Research Methods, Instruments, & Computers*, *36*(3), 402–407.

Oldfield, R. C., & Wingfield, A. (1964). The time it takes to name an object. *Nature*, *202*, 1031–1032.

Paivio, A. (2013). Dual coding theory, word abstractness, and emotion: A critical review of kousta et al. (2011). *Journal of Experimental Psychology: General*, *142*(1), 282–287. https://doi.org/10.1037/a0027004.

Paivio, A., & Begg, I. (1971). Imagery and comprehension latencies as a function of sentence concreteness and structure. *Perception & Psychophysics*, *10*(6), 408–412. https://doi.org/10.3758/BF03210323.

Roediger, H. L., Watson, J. M., McDermott, K. B., & Gallo, D. A. (2001). Factors that determine false recall: a multiple regression analysis. *Psychonomic Bulletin & Review*, *8*(3), 385–407. https://doi.org/10.3758/BF03196177.

Snodgrass, J. G., & Vanderwart, M. (1980). A standardized set of 260 pictures: Norms for name agreement, image agreement, familiarity, and visual complexity. *Journal of Experimental Psychology: Human Learning & Memory*, *6*(2), 174–215. https://doi.org/10.1037/0278-7393.6.2.174.

van Heuven, W. J. B., Mandera, P., Keuleers, E., & Brysbaert, M. (2014). SUBTLEX-UK: A new and improved word frequency database for British English. *Quarterly Journal of Experimental Psychology*, *67*(6), 1176–1190. https://doi.org/10.1080/17470218.2013.850521.